Maximum Likelihood Estimation and Regression

Parthiban Rajendran
parthi292929@gmail.com

October 25, 2018

Contents

Chapter 1

Maximum Likelihood Estimation

1.1 Introduction

Encrypted Introduction as in textbooks

Suppose that we have a random variable X, whose pdf or pmf is known but the distribution depends on an unknown parameter, say θ, that may have any value in a **parameter space** Ω. For instance, it might be $f(x;\theta) = \theta^2(1-\theta)^{1-x}$, $0 < x < \infty$ and $\theta \epsilon \Omega$. In certain instances, an experiments needs **to select one member** of the entire possibilities of θ family, $\{f(x;\theta), \theta \epsilon \Omega\}$. That is, he needs a **point estimate** $\hat{\theta}$, the value of the parameter that corresponds to selected pdf or pmf.

One of the most common estimation scenario is to take a random sample set from the selected distribution (pdf or pmf) and try to estimate θ of the distribution. That is, we repeat the experiment to take m number of samples, observe the sample $X_1, X_2, \cdots X_m$, and try to estimate θ by using observations $x_1, x_2, \cdots, x_m$

The function we will use to estimate the θ, is called, **estimator**, $u(X_1, X_2, \cdots, X_n)$, and we represent the computed **estimate** as $u(x_1, x_2, \cdots, x_m)$. Our expectation is, this estimate should be as close to θ as possible. Since we are estimating only one of all possible $\theta \ \epsilon \ \Omega$, $u(x_1, x_2, \cdots, x_m)$ is called a **point estimator**

Decrypting it

Suppose we flip a coin. The outcome of how many heads we get, is a Bernoulli distribution. Let us describe it with random variable X, that is, X denotes number of heads in an outcome and since we flip only once, its values are $X = 0, 1$. That is, getting no heads or 1 head. We do know its pmf as $f(x;p) = p^2(1-p)^{1-x}$, $0 < x < \infty$ and $0 \leq p \leq 1$. Note, p here is the mystic θ we just talked about, and the **parameter space** Ω is from 0 to 1. **There is always one p associated with any Bernoulli distribution** which we need to find out of all possibilities between 0 and 1 inclusive. The given Bernoulli distribution depends on this p and we set out to find that out one p value. That is, we need a **point estimate** $\hat{p}$, the value of the paramter that corresponds to selected Bernoulli distribution.

One of the most common estimation scenario is to take a random sample set from the Bernoulli distribution and try to estimate p of the distribution. That is, we flip the coin to take m number of samples, observe the sample $X_1, X_2, \cdots X_m$, and try to estimate p by using observations $x_1, x_2, \cdots, x_m$. The observations might be something like $1, 0, 0, 1, 0, 1, \cdots, 1$, 1 indicating heads and tails otherwise.

The function we will use to estimate the p is called **estimator** $u(X_1, X_2, \cdots, X_n)$ and we represent the computed **estimate** as $\hat{p} = u(x_1, x_2, \cdots, x_m)$. Our expectation is $\hat{p}$ should be as close to real p as possible. Since we are estimating only one $\hat{p}$ of all possible $[0, 1]$ range, $\hat{p} = u(x_1, x_2, \cdots, x_m)$ is called a **point estimator**

1.2 Bernoulli Distribution

1.2.1 Theory

Suppose we flip a coin, **only once**. Then,

Probability Mass Function

$$f(x; p) = P(X = x) = p^x(1-p)^{1-x} \qquad x = 0, 1$$

Mean

$$\overline{X} = E[X] = \sum_{k=0}^{1} X_k \cdot p(X_k) = p$$

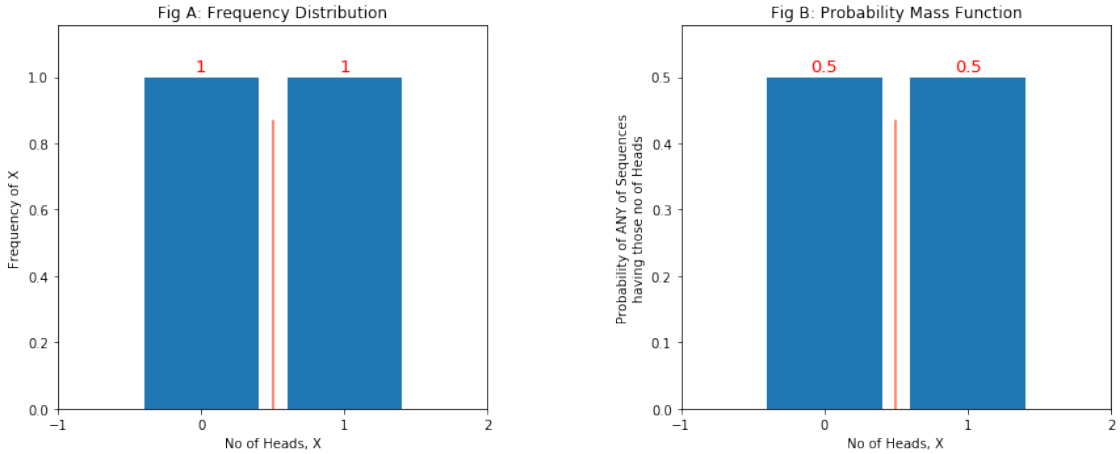
Variance

$$\sigma^2 = Var(X) = \sum_{k=0}^{1} (X_k - \bar{X})^2 p(X_k) = E(X^2) - [E(X)]^2 = p(1-p)$$

1.2.2 Example: Fair coin

Let us flip a **fair coin**, so we know $p = 0.5$. If X is a random variable indicating no of heads in the final outcome, then the probability mass function of X, would be as below.

```
The mean:0.5
```



Probability Mass Function for X=1

$f(x; p) = P(X = x) = p^x q^{n-x} = (0.5)^1(0.5)^{1-1} = 0.5$

Mean

$$\overline{X} = E[X] = \sum_{k=0}^{1} X_k \cdot p(X_k) = p = 0.5$$

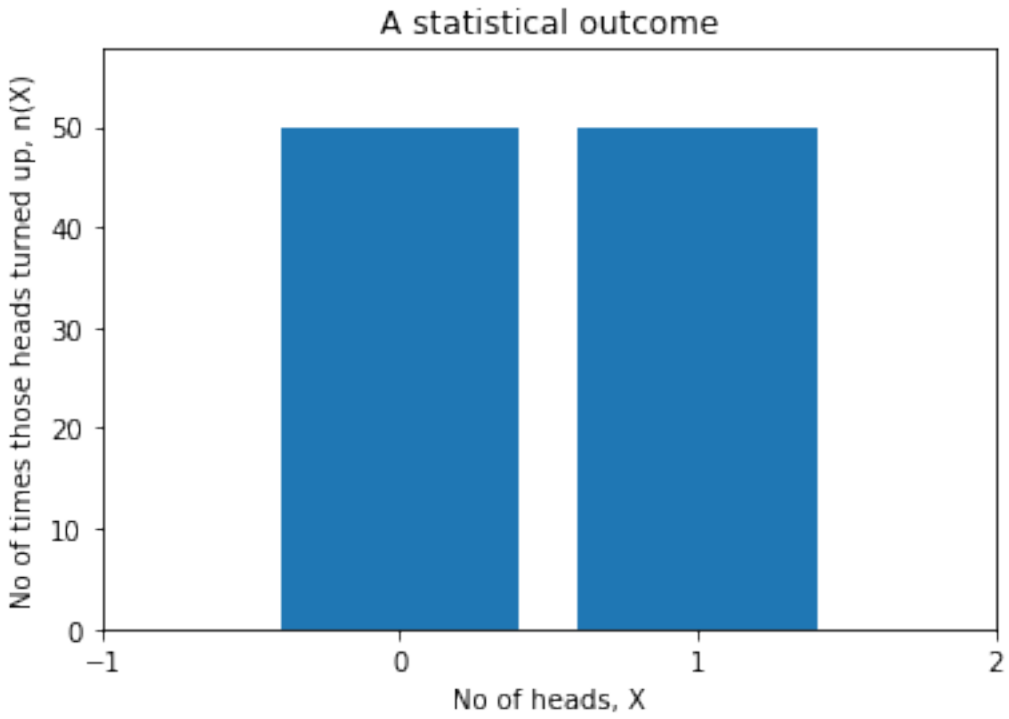
Variance

$$\sigma^2 = Var(X) = \sum_{k=0}^{1} (X_k - \bar{X})^2 p(X_k) = E(X^2) - [E(X)]^2 = p(1-p) = (0.5)(0.5) = 0.25$$

Statistical Outcome

Above example was for a fair coin, so $p = 0.5$, but p could have varied anywhere between 0 and 1 in reality (that is, coin might be loaded). $0 \leq p \leq 1$. So the question is if we observe a set of samples $X_1, X_2, \cdots, X_m$, with values $x_1, x_2, \cdots, x_m$ how do we determine the best value for p? We need a statistical procedure to determine the maximum likelihood of value p, given $X_1, X_2, \cdots, X_m$.

Suppose we have conducted such an experiment and turns out below is the frequency distribution of the outcome (Note this is similar to Fig A we just saw above). It reads, we got 50 heads and 50 tails out of $m = 100$ trails.



One could then *estimate* the underlying p as simply the mean value as below. If $X_1, X_2, \cdots, X_m$ are the total m number of samples, then

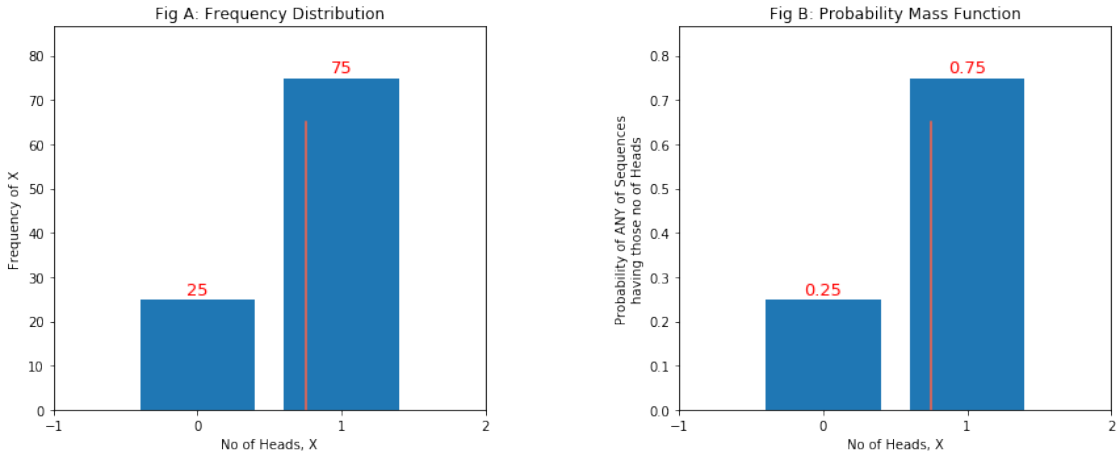
$$\hat{p} = \frac{\sum_{i=0}^{1} X_i n(X_i)}{m} = \frac{0(50) + 1(50)}{50 + 50} = 0.5$$

Thus the point estimator $\hat{p}$ for p from given sample set is 0.5. Since we already know the theoretical p of fair coin, we are sure how best is our estimation $\hat{p}$. This function which we just used is the maximum likelihood estimation function.

1.2.3 Example: Loaded coin

Suppose we have a **loaded coin**, and assume, we hypotheticall know, $p = 0.75$. If X is a random variable indicating no of heads in the final outcome, then the probability mass function of X, would be as below.

```
The mean:0.75
```



Probability Mass Function for X=1

$f(x;p) = P(X = x) = p^x q^{n-x} = (0.75)^1(0.75)^{1-1} = 0.75$

Mean

$\overline{X} = E[X] = \sum_{k=0}^{1} X_k \cdot p(X_k) = 0(0.25) + 1(0.75) = 0.75 = p$

Variance

$\sigma^2 = Var(X) = \sum_{k=0}^{1} (X_k - \bar{X})^2 p(X_k) = E(X^2) - [E(X)]^2 = p(1-p) = (0.75)(0.25) = 0.1875$

Statistical Outcome

You see, Fig A above actually represents how a frequency distribution of an experiment would be, provided the coin was loaded. So we could directly calculate the estimate as below.

$$\hat{p} = \frac{\sum_{i=0}^{1} X_i n(X_i)}{m} = \frac{0(25) + 1(75)}{25 + 75} = 0.75$$

Thus, from the sample observations $x_1, x_2, .., x_m$, we are able to calculate $\hat{p}$. Thus, for Bernoulli distributions, the mean $\overline{X}$ of the sample set is the MLE of p.

Bernoulli Distribution; m trials

$$\hat{p} = u(x_1, x_2, \cdots, x_m) = \frac{\sum_{i=0}^{1} X_i n(X_i)}{m} \rightarrow p \tag{1.1}$$

1.2.4 MLE Derivation

Establishing the likelihood function

We are just empirically convinced that the **estimator** $u(X_1, X_2, \cdots, X_m) = \frac{\sum_{i=0}^{1} X_i n(X_i)}{m}$ gives us maximum likelihood value $\hat{p}$ of p. Here we try to prove mathematically that is the best estimator indeed out of all possibilities of p for given sample set.

We start with the sample set. Remember, that is all we have to look at and try backwards to find maximum likelihood of p that would have resulted in that sample set. Let us take our loaded coin example above. Out of $m = 100$ trials, we got about 75 as heads(1) and 25 as tails(0). Our entire sample set might look like this: $\{1, 0, 1, 1, 0, \cdots, 1\}$ with length of m.

We will try to figure out the probability mass function of this sample result and see, when that *joint pmf* maxes out for a given p. Why? because, that is the best case, where we would have gotten all these values one after another. Any other **joint pmf** value, would have given lesser probability for all these values to occur simultaneously.

Let me break it down. Here we are wondering *I have this series of outcomes* and I need to find the probability of this occurance. This is a joint occurance, thus we would find the joint probability. Remember, each trial is independent.

Suppose you have events A and B, and then asked, what is the probability of both A and B happeningn, that is $A \cap B$, then you would say, $p(A, B) = p(A \cap B) = p(A)p(B)$. Similarly, for $\{1, 0, 1, 1, 0, \cdots, 1\}$,

$$p(X_1 = 1, X_2 = 0, X_3 = 1, \cdots, X_m = 1) = p(X_1 = 1)p(X_2 = 0)p(X_3 = 1) \cdots p(X_m = 1)$$

Generalizing for any sample set,

$$p(X_1 = x_1, X_2 = x_2, X_3 = x_3, \cdots, X_m = x_m) = p(X_1 = x_1)p(X_2 = x_2)p(X_3 = x_3) \cdots p(X_m = x_m)$$

We already know,

$$p(X_i = x_i) = f(x_i; p) = p^{x_i}(1 - p)^{1-x_i}, \quad x_i = 0, 1 \quad 0 \leq p \leq 1$$

Combining above two equations, we get,

$$\begin{aligned} p(X_1 = x_1, X_2 = x_2, X_3 = x_3, \cdots, X_m = x_m) &= \prod_{i=1}^{m} p^{x_i}(1 - p)^{1-x_i} \\ &= p^{x_1+x_2+\cdots+x_m}(1 - p)^{(1_1+1_2+1_3\cdots+1_m)-(x_1+x_2+\cdots+x_m)} \\ &= p^y(1 - p)^{m-y}, \quad \text{where } y = \sum_{i=1}^{m} x_i \end{aligned} \tag{1.2}$$

Now, given the sample set, we have arrived at a *joint pmf* function of p. This is called the **likelihood function**. Let us denote it by $L(p)$. So for a Bernoulli distribution we just established the likelihood function as

Bernoulli Distribution; m trials

The likelihood function,

$$L(p) = p^y(1-p)^{m-y}, \quad \text{where } y = \sum_{i=1}^{m} x_i \quad 0 \le p \le 1 \tag{1.3}$$

Establishing the p range

We now need to find out, what is the p value for which, we would get $L(p)$ to max out. Why? Recall, that is the maximum probability our combination of sample values would have occurred. That is the best value we could find. (eh, as long as there is only one peak or maximum for $L(p)$, but that is another problem for another time).

Let us check out what happens to $L(p)$ for edge values.

Case 1: Getting all tails in sample set

This means,

$$y = \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} 0 = 0 \therefore L(p) = p^0(1-p)^{m-0} = (1-p)^m$$

The above function $(1-p)^m$ will have its maximum when $p = 0$. Logically for any $p > 0$, $1 - p$ would be lesser. We got all tails, that means, the probability of getting heads should be 0 which also makes sense (for other values of p, there is still a chance to get all tails, but comparitively lesser chance. $p = 0$ has the maximum probability of getting us all tails, so that is our best estimate in this case)

$$\therefore \text{when } \; y = 0, \quad L_{max}(p) = 1 \quad \implies \quad \hat{p} = 0$$

Case 2: Getting all heads in sample set

This means

$$y = \sum_{i=1}^{m} x_i = \sum_{i=1}^{m} 1 = m \therefore L(p) = p^m(1-p)^{m-m} = p^m$$

The above function p^m will have its maximum value when $p = 1$ because, the maximum value of p is 1. So maximum of p^m is also 1. Any $p < 1$ will result in reduced p^m also accordingly.

$$\therefore \text{when } \; y = m, \quad L_{max}(p) = 1 \quad \implies \quad \hat{p} = 1$$

Case 3: Neither Case 1 or Case 2

This means, y is neither 0 nor m. That is, $0 < y < m$

Case 3.1: When $\hat{p} = 0$

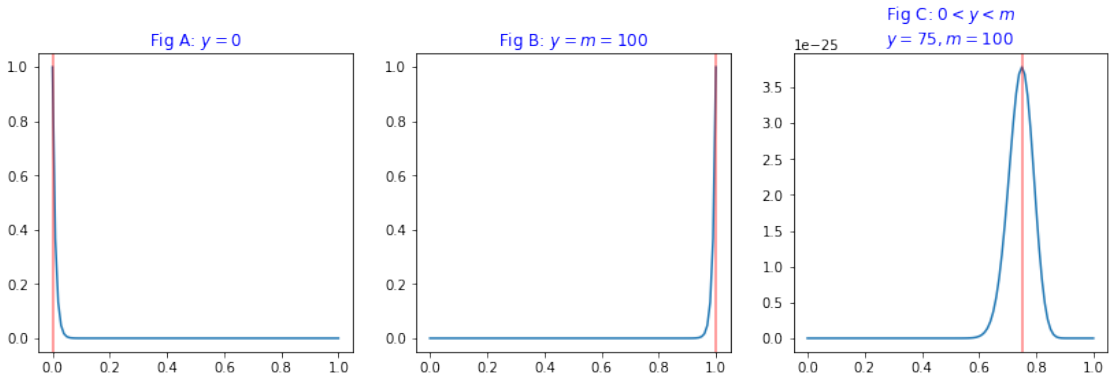
$$L(0) = 0^y(1-0)^{m-y} = 0$$

Case 3.2: When $\hat{p} = 1$

$$L(1) = 1^y(1-1)^{m-y} = 1^y 0^{m-y} = 0$$

In fact, we could already check for any sample set of $m = 100$, how the function $L(p)$ behaves. Only by observing where the function reaches maximum,

1. In Fig A, when $y = 0$, potential candidate for $\hat{p}$ is 0 because that is where $L(p)$ reaches maximum of 1, and then quickly faded to 0 thereafter.
2. In Fig B, when $y = m$, potential candidate for $\hat{p}$ is 1 because that is where $L(p)$ reaches maximum of 1, and was 0 till then.
3. In Fig C, when $0 < y < m$, potential candidate for $\hat{p}$ is y/m because that is where $L(p)$ reaches maximum (though not 1 but dependent on p, y). Note, at $p = 0$ and $p = 1$, $L(p)$ is 0 already. In fact, it is 0 for most of p which is an interesting insight. It is only when we near the y/m the $L(p)$ rises and falls.



Bernoulli Distribution; m trials

- When $y = 0$, we already know best estimate as $\hat{p} = 0$ as $L(p)$ reaches maximum value 1.
- When $y = m$, we already know best estimate as $\hat{p} = 1$ as $L(p)$ reaches maximum value 1.
- When $0 < y < m$, we already know $L(p)$ reaches minimum value 0, when $\hat{p} = 0$ or $\hat{p} = 1$. Since we set out to find the $\hat{p}$ at which $L(p)$ reaches maximum, we could ignore $\hat{p} = 0$ and $\hat{p} = 1$ for $0 < y < m$
- Combining above points, we could say, we need to focus only on cases $0 < y < m$, and in that, only where $0 < p < 1$, because that is where $L(p)$ attains maximum for which we need to find respective $\hat{p}$. In other words, if you get all tails or all heads in a sample set, you know your best estimate $\hat{p}$ already. We set out to find for rest of the cases where $0 < p < 1$

Finding the optimal estimate

We just saw via graph (Fig C), the nature of $L(p)$ where we could once again convinced of the optimality of y/m as best candidate for point estimator $\hat{p}$, however we also could and should prove mathematically that is the case. From calculus, we know that the derivative $\frac{dL(p)}{dp}$ will be 0 when $L(p)$ reaches maximum (one could refer to appendix 3.2 for a quick recap on this concept). So by taking the derivative and equating to 0, we could derive the optimal p.

$$L(p) = p^y(1-p)^{m-y}$$

$$\frac{dL(p)}{dp} = \frac{d\{p^y(1-p)^{m-y}\}}{p}$$

From product_rule of derivatives, in Leibniz's notation,

$$\frac{d(u.v)}{dx} = \frac{du}{dx}.v + u.\frac{dv}{dx}$$

So,

$$\frac{dL(p)}{dp} = \frac{d(p^y)}{dp}.(1-p)^{m-y} + \frac{d\{(1-p)^{m-y}\}}{dp}.p^y$$

The term $\frac{d\{(1-p)^{m-y}\}}{dp}$ is little tricky.
Let $u = (1-p)$, $k = (m-y)$, then by using chain rule

$$\begin{aligned}\frac{d\{(1-p)^{m-y}\}}{dp} = \frac{d\{u^k\}}{dp} = \frac{\partial u^k}{\partial u}\frac{\partial u}{\partial p} = ku^{k-1}\frac{\partial(1-p)}{\partial p} \\ = ku^{k-1}(-1) = -ku^{k-1} \\ = -(m-y)(1-p)^{m-y-1}\end{aligned}$$

Substituting,

$$\begin{aligned}\frac{dL(p)}{dp} = yp^{y-1}.(1-p)^{m-y} - (m-y)(1-p)^{m-y-1}.p^y \\ = yp^yp^{-1}.(1-p)^{m-y} - (m-y)(1-p)^{m-y}(1-p)^{-1}.p^y \\ = p^y(1-p)^{m-y}\Big(yp^{-1} - (m-y)(1-p)^{-1}\Big) \\ = p^y(1-p)^{m-y}\Big(\frac{y}{p} - \frac{m-y}{1-p}\Big)\end{aligned}$$

$\therefore$

$$\frac{dL(p)}{dp} = L(p)\Big(\frac{y}{p} - \frac{m-y}{1-p}\Big)$$

Equating it to 0, and noting that, when the derivative is 0, $L(p)$ reaches maximum, so it cannot be 0, we get,

$$\begin{aligned}\frac{dL(p)}{dp} &= L(p)\Big(\frac{y}{p} - \frac{m-y}{1-p}\Big) = 0 \\ &\implies \Big(\frac{y}{p} - \frac{m-y}{1-p}\Big) = 0 \\ &= y - yp - mp + yp = 0 \\ &= y - mp = 0 \implies y = mp\end{aligned}$$

$\therefore$, at $\dfrac{dL(p)}{dp} = 0$, $L(p)$ reaches maximum at $y = mp$ or equivalently, in terms of p, $p = y/m$. Replacing y with its original summation $\sum_{i=1}^{m} x_i$, we finally get,

Bernoulli Distribution; m trials

$$\hat{p} = \frac{y}{m} = \frac{\sum_{i=1}^{m} x_i}{m} = \overline{x} \to p \tag{1.4}$$

Thus, proved. Note it would have been easier to prove with taking logarithm on both sides of $L(p)$. Also note this is same as 1.1 except that, earlier we used frequency distribution's data so formula looks slightly different.

1.3 Binomial Distribution

Let us try similar approach for a Binomial distribution. Remember, binomial distribution is simply Bernoulli distributions repeated. When each Bernoulli event is independent, the resultant combined distribution (or the associated pmf) would be a Binomial distribution

1.3.1 Theory

Suppose we flip a coin, n no of times. Then,

Probability Mass Function

$$f(x;p) = P(X = x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \qquad x = 0, 1, \cdots, n$$

Mean

$$\overline{X} = E[X] = \sum_{k=0}^{n} X_k \cdot p(X_k) = np$$

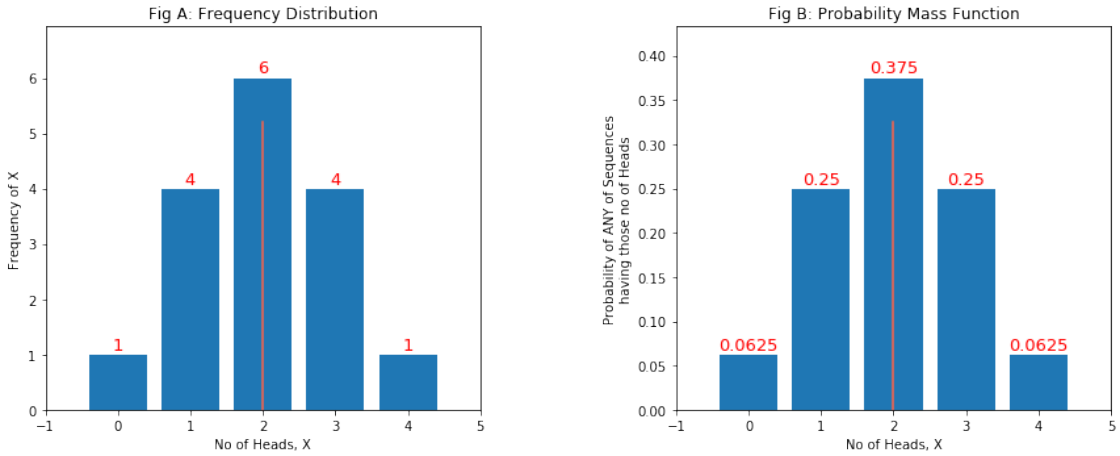
Variance

$$\sigma^2 = Var(X) = \sum_{k=0}^{n} (X_k - \bar{X})^2 p(X_k) = E(X^2) - [E(X)]^2 = np(1-p)$$

1.3.2 Example: Fair coin

Let us flip a **fair coin** (so we know $p = 0.5$), $n = 4$ times. If X is a random variable indicating no of heads in the final outcome, then the probability mass function of X, for $n = 4$ would be as below.

```
The mean:2.0
```



Probability Mass Function for X=2

$$f(x;p) = P(X = x) = \frac{n!}{x!(n-x)!}p^x q^{n-x} = \frac{4!}{2!(4-2)!}(0.5)^2(0.5)^{4-2} = 0.375$$

Mean

$$\overline{X} = E[X] = \sum_{k=0}^{n} X_k \cdot p(X_k) = np = 4(0.5) = 2$$

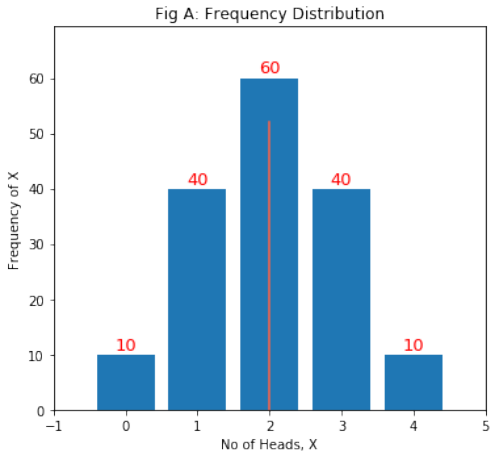
Variance

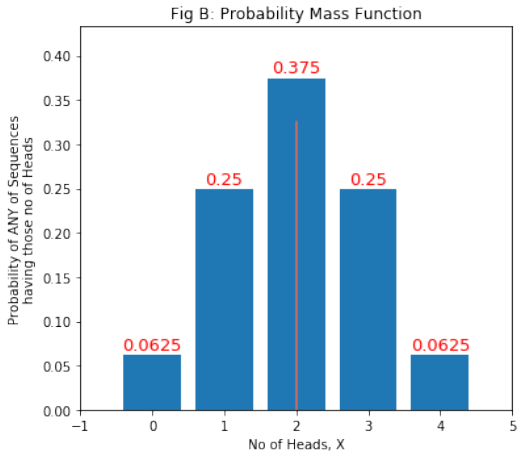
$$\sigma^2 = Var(X) = \sum_{k=0}^{n} (X_k - \bar{X})^2 p(X_k) = E(X^2) - [E(X)]^2 = npq = 4(0.5)(0.5) = 1$$

Statistical Outcome

Suppose we conduct an experiment of flipping the fair coin $n = 4$ times and observe the result $x_1 = \{0, 0, 1, 0\}$ this would mean, $TTHT$ or $X = 1$ heads. So frequency of (X=1) adds by 1. Similarly, we repeat the experiment $m = 160$ times (just for convenience of numbers which you will realize shortly). So our samples would be $X_1, X_2, \cdots, X_m$. We note down the frequency of $X = x$ in each experiment and plot the graph. Suppose we get a discrete frequency distribution graph as below (left one and we could derive right one from frequency data).

```
The mean:2.0
```





Let us calculate the mean value of above frequency distribution, $\overline{X}$.

$$\overline{X} = \frac{\sum_{i=0}^{n} X_i n(X_i)}{m} = \frac{0(10) + 1(40) + 2(60) + 3(40) + 4(10)}{10 + 40 + 60 + 40 + 10} = \frac{320}{160} = 2$$

Wait a minute, our theoretical p was 0.5?! Yes, in case of Binomial, we go further as dividing the mean $\overline{X}$ by no of flips n.

$$\hat{p} = \frac{\overline{X}}{n} = \frac{\sum_{i=0}^{n} X_i n(X_i)}{nm} = \frac{2}{4} = 0.5$$

It happens that, in case of binomial distribution, the best estimator $\hat{p}$ for p would be $\overline{X}/n$. Thus, from the sample observations $x_1, x_2, .., x_m$, we are able to calculate $\hat{p}$.

Binomial Distribution; n flips; m trials

$$\hat{p} = u(x_1, x_2, \cdots, x_m) = \frac{\sum_{i=0}^{n} X_i n(X_i)}{nm} \rightarrow p \tag{1.5}$$

Note: By substituting, $n = 1$ in above equation we get the MLE for Bernoulli as expected. We now need to wonder how to prove, if this is the best MLE.

1.3.3 MLE Derivation

Let n be the number of flips and m be the number of trials. Then our sample set could be looking something like this:

$$X_1 = \{0, 1, ...x_{1n}\} = 1 \text{ heads}$$
$$X_2 = \{1, 1, ...x_{2n}\} = 4 \text{ heads}$$
$$\cdots$$
$$\cdots$$
$$X_m = \{1, 0, ...x_{mn}\} = 2 \text{ heads}$$

Generalizing,

$$
\begin{aligned}
X_1 &= \{x_{11}, x_{12}, ...x_{1n}\} = x_1 \text{ heads} \\
X_2 &= \{x_{21}, x_{22}, ...x_{2n}\} = x_2 \text{ heads} \\
&\cdots \\
&\cdots \\
X_m &= \{x_{m1}, x_{m2}, ...x_{mn}\} = x_m \text{ heads}
\end{aligned}
$$

We already know, for a single Bernoulli distribution, the pmf is given by

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Therefore, *given* an observed sample set X, the combined or joint probability of all sample observations in the sample set could be given by, as likelihood function

$$
\begin{aligned}
L(p) = P(X_1 = x_1; X_2 = x_2; \cdots X_m = x_m) &= P(X_1 = x_1)P(X_2 = x_2)\cdots P(X_m = x_m) \\
&= \prod_{i=1}^{m} P(X_i = x_i) \\
&= \prod_{i=1}^{m} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} \qquad (1.6) \\
&= \left\{ \prod_{i=1}^{m} \binom{n}{x_i} \right\} \left\{ \prod_{i=1}^{m} p^{x_i}(1-p)^{n-x_i} \right\}
\end{aligned}
$$

This time we will use natural logarithms to find the maximum. Recalling product rule of natural logarithms 3.7

$$
\begin{aligned}
\left\{ \prod_{i=1}^{m} \binom{n}{x_i} \right\} &= \left\{ \binom{n}{x_1} \binom{n}{x_2} \cdots \binom{n}{x_m} \right\} \\
\implies ln\left\{ \prod_{i=1}^{m} \binom{n}{x_i} \right\} &= \left\{ ln\binom{n}{x_1} + ln\binom{n}{x_2} + \cdots + ln\binom{n}{x_m} \right\} \qquad (1.7) \\
&= \sum_{i=1}^{m} ln\binom{n}{x_i}
\end{aligned}
$$

$$\left\{ \prod_{i=1}^{m} p^{x_i}(1-p)^{n-x_i} \right\} = p^{(x_1+x_2+\cdots+x_m)}(1-p)^{(n_1+n_2+\cdots+n_m)-(x_1+x_2+\cdots+x_m)}$$

Let $y = \sum_{i=1}^{m} x_i$. The preceding equation could be written as,

$$\left\{ \prod_{i=1}^{m} p^{x_i}(1-p)^{n-x_i} \right\} = p^y (1-p)^{mn-y}$$

Taking natural logrithm on both sides and using product rule,

$$\begin{aligned} ln\Big\{\prod_{i=1}^{m} p^{x_i}(1-p)^{n-x_i}\Big\} &= ln\Big\{p^{y}(1-p)^{mn-y}\Big\} \\ &= y\Big\{ln(p)\Big\} + (mn-y)\Big\{ln(1-p)\Big\} \end{aligned} \tag{1.8}$$

Using 1.8 and 1.7 in 1.6, and again using product rule,

$$ln\Big\{L(p)\Big\} = \sum_{i=1}^{m} ln\binom{n}{x_i} + y\Big\{ln(p)\Big\} + (mn-y)\Big\{ln(1-p)\Big\}$$

To find the maximum, let us equate the derivative of $ln\Big\{L(p)\Big\}$ w.r.t p to 0.

$$\frac{d\Big\{\{lnL(p)\}\Big\}}{dp} = 0$$

$$\implies \frac{d\Big\{\sum_{i=1}^{m} ln\binom{n}{x_i}\Big\}}{dp} + \frac{d\Big\{y\{ln(p)\}\Big\}}{dp} + \frac{d\Big\{(mn-y)\{ln(1-p)\}\Big\}}{dp} = 0$$

The first term has no p, so the derivative w.r.t p becomes 0. And for rest of components, by referring to derivatives of natural logarithms 3.9 and 3.10,

$$\begin{aligned} 0 + y\frac{d\{ln(p)\}}{dp} + (mn-y)\frac{d\{ln(1-p)\}}{dp} &= 0 \\ y\frac{1}{p} + (mn-y)\frac{-1}{1-p} &= 0 \\ \implies \frac{y}{p} &= \frac{mn-y}{1-p} \\ y - yp &= mnp - yp \\ y &= mnp \\ p &= \frac{y}{mn} \end{aligned}$$

Substituting, $y = \sum_{i=1}^{m} x_i$, we get, $p = \frac{\sum_{i=1}^{m} x_i}{mn}$. Note that this is consistent with our empirical evidence 1.5

Binomial Distribution; n flips; m trials

$$\hat{p} = \frac{y}{mn} = \frac{\sum_{i=1}^{m} x_i}{mn} = \frac{\overline{x}}{n} \to p \tag{1.9}$$

1.4 Normal Distribution

Let us try Normal distribution as an example of MLE for continuous distribution.

1.4.1 Theory

Probability Density Function

$$f(x;p) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big[-\frac{(x_i-\mu)^2}{2\sigma^2}\Big]$$

Mean

μ

Variance

σ^2

1.4.2 MLE Derivation

Let $X_1, X_2, \cdots, X_m$ be a random sample from $N(\theta_1, \theta_2)$, where both the parameters belong to parameter space defined as

$$\Omega = \{(\theta_1, \theta_2) : -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\}$$

Letting $\theta_1 = \mu, \theta_2 = \sigma^2$, one might be tempted to attempt the likelihood function as below (as combined *joint probability* of getting all the sample data.

$$L(\theta_1, \theta_2) = P(X_1 = x_1; X_2 = x_2; \cdots; X_n = x_m)$$

However, unlike a *pmf* which directly gives $P(X_i = x_i)$, a *pdf* only a function and always needs integration to find the probability area. That is, if x_1 is a sample observation from $N(\theta_1, \theta_2)$, then $P(X_1 = x_1) = 0$, and we are not interested in that in particular (which was a wrong notion implicitly implanted while attempting joint *pmf*). Instead we are interested in a collective probability density *function* of all samples' individual probability densities.

That is, below is a *continuous* pdf for sample X_1

$$A = f(x_1; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}}\exp\Big[-\frac{(x_1-\theta_1)^2}{2\theta_2}\Big]$$

But when we want to find a probability with above *pdf* its always in a range. For example,

$$P(X_1 \leq a) = \int_{-\infty}^{a} f(x_1; \theta_1, \theta_2)dx_1 \tag{1.10}$$

Similarly, for another sample X_2 from same *pdf*,

$$B = f(x_2; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}}\exp\Big[-\frac{(x_2-\theta_1)^2}{2\theta_2}\Big]$$

And for that, for an interesting range, the probability could be something like below.

$$P(X_2 \leq b) = \int_{-\infty}^{b} f(x_2; \theta_1, \theta_2)dx_2 \tag{1.11}$$

Note A and B are the *functions* while, eq. 4 and 6 denote a probability calculated out of those functions. When we say, we are interested in *joint pdf*, we are interested in the multiplication of

the *functions* A and B (because they are independent), and not probabilities like 1.10 and 1.11. The probability of any joint interested event could be calculated in resultant function AB. That is,

$$AB = f(x_1, x_2; \theta_1, \theta_2) = \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi\theta_2}} \exp\Big[- \frac{(x_i - \theta_1)^2}{2\theta_2} \Big]$$

And then, in this *joint pdf* I could calculate interested probabilities, for example,

$$P(X_1 \leq a; X_2 \leq b) = \int_{-\infty}^{x_1=a} \int_{-\infty}^{x_2=b} \prod_{i=1}^{2} \frac{1}{\sqrt{2\pi\theta_2}} \exp\Big[- \frac{(x_i - \theta_1)^2}{2\theta_2} \Big]$$

Generalizing,

$$P(X_1 \leq x_1; X_2 \leq x_2) = \prod_{i=1}^{2} \int_{-\infty}^{x_i} \frac{1}{\sqrt{2\pi\theta_2}} \exp\Big[- \frac{(x_i - \theta_1)^2}{2\theta_2} \Big]$$

Not just left area, but any probability of interest could be calculated after this step. For example,

$$P(X_1 \geq x_1; X_2 \geq x_2) = \prod_{i=1}^{2} \int_{x_i}^{\infty} \frac{1}{\sqrt{2\pi\theta_2}} \exp\Big[- \frac{(x_i - \theta_1)^2}{2\theta_2} \Big]$$

This is why, unlike *pmf*, for a pdf,

$$\begin{aligned} f(x_1, x_2; \theta_1, \theta_2) &= f(x_1; \theta_1, \theta_2) f(x_2; \theta_1, \theta_2) \\ &\neq P(X_1 \leq x_1; X_2 \leq x_2) \\ &\neq P(X_1 \geq x_1; X_2 \geq x_2) \\ &\neq P(X_1 = x_1; X_2 = x_2) \end{aligned}$$

Thus, a better notion of *joint pdf* as the likelihood function is

$$\begin{aligned} L(\theta_1, \theta_2) = f(x_1, x_2, \cdots, x_m; \theta_1, \theta_2) &= f(x_1; \theta_1, \theta_2) f(x_2; \theta_1, \theta_2) \cdots f(x_m; \theta_1, \theta_2) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\theta_2}} \exp\Big[- \frac{(x_i - \theta_1)^2}{2\theta_2} \Big] \end{aligned}$$

Taking natural logarithms on both sides,

$$\begin{aligned} ln\ L(\theta_1, \theta_2) &= ln \Bigg\{ \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\theta_2}} \exp\Big[- \frac{(x_i - \theta_1)^2}{2\theta_2} \Big] \Bigg\} \\ &= ln \Bigg\{ \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\theta_2}} \Bigg\} + ln \Bigg\{ \prod_{i=1}^{m} \exp\Big[- \frac{(x_i - \theta_1)^2}{2\theta_2} \Big] \Bigg\} \end{aligned} \tag{1.12}$$

Note that the term,

$$\begin{aligned} ln \Bigg\{ \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\theta_2}} \Bigg\} &= ln\ \left(\frac{1}{\sqrt{2\pi\theta_2}} \right)^m \\ &= ln\ (2\pi\theta_2)^{\frac{-m}{2}} \\ &= \left(\frac{-m}{2} \right) ln\ (2\pi\theta_2) \end{aligned}$$

And for the 2nd term,

$$ln\Big\{\prod_{i=1}^{m}\exp\Big[-\frac{(x_i-\theta_1)^2}{2\theta_2}\Big]\Big\} = ln\Big\{\exp\sum_{i=1}^{m}\Big[-\frac{(x_i-\theta_1)^2}{2\theta_2}\Big] = ln\Big\{\exp\Big[-\frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2}\Big]\Big\}$$

Recall that,
$a = e^{ln(a)}$ so when $a = e$, $e = e^{ln(e)} \implies ln(e) = 1$. Applying that,

$$ln\Big\{\prod_{i=1}^{m}\exp\Big[-\frac{(x_i-\theta_1)^2}{2\theta_2}\Big]\Big\} = \Big[-\frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2}\Big]$$

Applying above derivations in eq 1.12,

$$ln\ L(\theta_1,\theta_2) = \Big(\frac{-m}{2}\Big)\ ln\ (2\pi\theta_2) + \Big[-\frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2}\Big] \tag{1.13}$$

In order to evaluate when $lnL(\theta_1,\theta_2)$ reaches maximum, let us take the partial derivatives w.r.t to θ_1,θ_2 and equate them to 0 (refer appendix 3.2.7)
Assuming θ_2 as a constant,

$$\begin{aligned}\frac{\partial L}{\partial\theta_1} &= 0 + \frac{\partial\Big[-\frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2}\Big]}{\partial\theta_1}\\ &= \Big[-\frac{\sum_{i=1}^{m}2(x_i-\theta_1)}{2\theta_2}\Big]\\ &= \Big[-\frac{\sum_{i=1}^{m}(x_i-\theta_1)}{\theta_2}\Big]\end{aligned} \tag{1.14}$$

Taking $\frac{\partial L}{\partial\theta_1} = 0$, we get,

$$\begin{aligned}\Big[-\frac{\sum_{i=1}^{m}(x_i-\theta_1)}{\theta_2}\Big] &= 0\\ \implies \sum_{i=1}^{m}(x_i-\theta_1) &= 0\\ \sum_{i=1}^{m}x_i - m\theta_1 &= 0\\ \implies \theta_1 = \frac{1}{m}\sum_{i=1}^{m}x_i &= \overline{x}\end{aligned} \tag{1.15}$$

Assuming θ_1 as constant,

$$\frac{\partial L}{\partial\theta_2} = \Big(\frac{-m}{2}\Big)\frac{\partial ln(2\pi\theta_2)}{\partial\theta_2} + \frac{\partial}{\partial\theta_2}\Big[-\frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2}\Big] \tag{1.16}$$

Taking the first term, recall 3.11 that $\frac{d(ln(cx))}{dx} = \frac{1}{x}$, yeah the constant disappears?!

$$\therefore \left(\frac{-m}{2}\right)\frac{\partial ln(2\pi\theta_2)}{\partial\theta_2} = \left(\frac{-m}{2}\right)\frac{1}{\theta_2}$$

Taking the second term,

$$\begin{aligned}
&\frac{\partial}{\partial\theta_2}\left[-\frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2}\right] \\
&= \frac{-\sum_{i=1}^{m}(x_i-\theta_1)^2}{2}\frac{\partial}{\partial\theta_2}\left(\frac{1}{\theta_2}\right) \\
&= \frac{-\sum_{i=1}^{m}(x_i-\theta_1)^2}{2}\left(\frac{\partial\theta_2^{-1}}{\partial\theta_2}\right) \\
&= \frac{-\sum_{i=1}^{m}(x_i-\theta_1)^2}{2}\left(-\theta_2^{-2}\right) \\
&= \frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2^2}
\end{aligned}$$

Substituting both in 1.16,

$$\frac{\partial L}{\partial\theta_2} = \left(\frac{-m}{2}\right)\frac{1}{\theta_2} + \frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2^2} \tag{1.17}$$

Taking $\frac{\partial L}{\partial\theta_2} = 0$, we get,

$$\begin{aligned}
&\left(\frac{-m}{2}\right)\frac{1}{\theta_2} + \frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2^2} = 0 \\
&\implies \frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{2\theta_2^2} = \left(\frac{m}{2}\right)\frac{1}{\theta_2}
\end{aligned}$$

Cancelling common terms on both sides,

$$\begin{aligned}
&\frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{\theta_2} = m \\
&\implies \theta_2 = \frac{\sum_{i=1}^{m}(x_i-\theta_1)^2}{m} \\
&\theta_2 = \frac{\sum_{i=1}^{m}(x_i-\overline{x})^2}{m}
\end{aligned} \tag{1.18}$$

1.4.3 Visualization

Visualizing the likelihood function $lnL(\theta_1,\theta_2)$, helps to comprehend the concept better especially when we observe where it maxes out to provide us the maximum likelihood estimators $\hat{\theta_1},\hat{\theta_2}$. Graphing the direct $L(\theta_1,\theta_2)$ is complicated, so we instead graphed the log likelihood function $ln(L(\theta_1,\theta_2))$. Sample set from a binomail distribution (which is approximated by normal distribution usually) is used to feed the $ln(L)$, and then its graph observed for varying values of θ_1,θ_2.

Sample setup

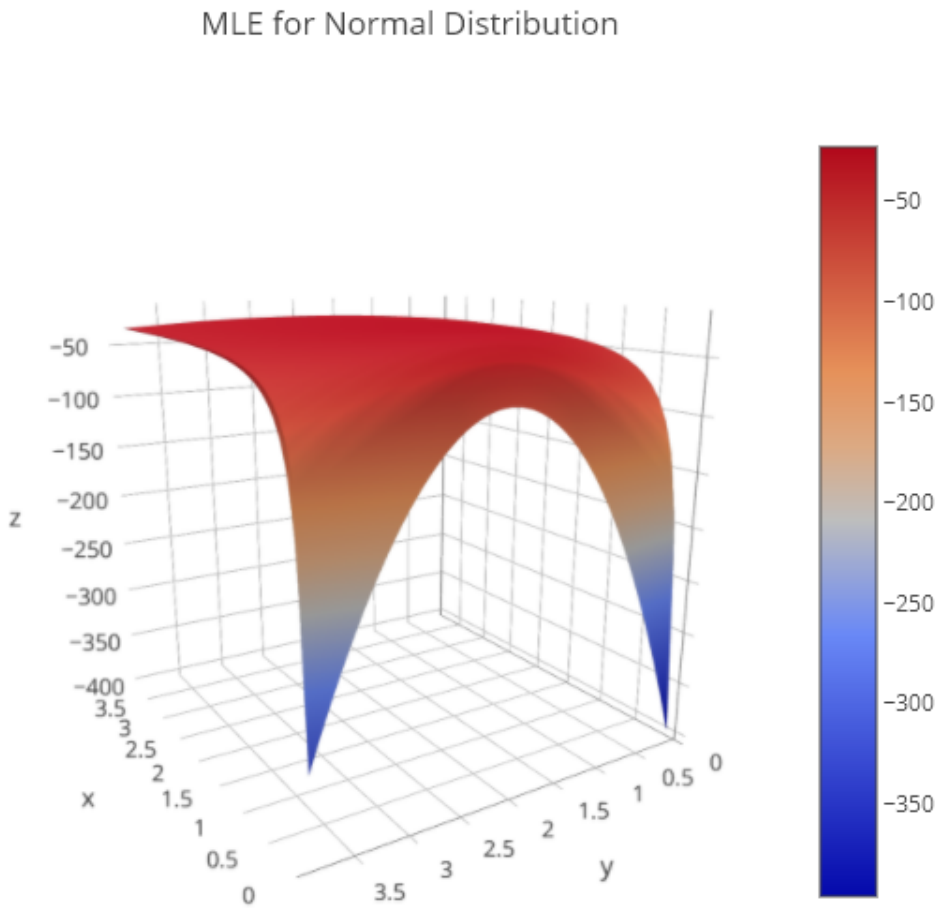
```
In[1]: x_i = [0,1,1,1,1,2,2,2,2,2,3,3,3,3,4]   # a binomial distribution
       # x_i = np.random.normal(2, 1.5, 100)  # one could also try this
       m = len(x_i)
       mean = sum(x_i)/m
       variance = sum([ (i-mean)**2 for i in x_i ])/m
       print('mean:{}, variance:{}'.format(mean, variance))
```

```
mean:2.0, variance:1.0
```

Graph

For brevity, graph code is hidden. I have created a separate interactive page which explains in detail, how the graph is created and also an interactive 3D view of below image at the end of it. Please check it out here.

Out[3]:



Finding the maximum from the computed $ln(L)$

The $ln(L)$ value was computed for different values of (θ_1, θ_2) along with fixed sum of sample sets as in the formula, to build the graph. Now, one could find the maximum value of that computed $ln(L)$, and note that, the respective (θ_1, θ_2) are indeed the mean and variance as we calculated in 1.15 and 1.18.

```
In[4]: df.loc[df['L'].idxmax()]
```

```
Out[4]: t1      2.000000
        t2      1.000000
        L     -22.703017
        Name: 789, dtype: float64
```

Normal Distribution

Thus, for any sample set from normal distribution with $N(\theta_1, \theta_2)$, Maximul Likelihood Estimators are

$$\hat{\theta_1} = \frac{1}{m}\sum_{i=1}^{m} x_i = \overline{x} \tag{1.19}$$

$$\hat{\theta_2} = \frac{\sum_{i=1}^{m}(x_i - \overline{x})^2}{m} \tag{1.20}$$

Chapter 2

Regression

2.1 The Simple Linear Regression Model

2.1.1 Introduction

Suppose we have a sample set (X, Y) of size m, that is $(X, Y) = \{(x_1, y_1), (x_2, y_2) \cdots (x_m, y_m)\}$. Then a simple linear model assumes a linear relationship between variables (x_i, y_i), and tries to estimate that. For example, observe a sample scatter plot of sample set in Figure 2.1. By looking at the figure, one could intuitively guess a linear relation between x and y variables as y increasing roughly with x. It is this we will try to find, and in that, find the best possible one.

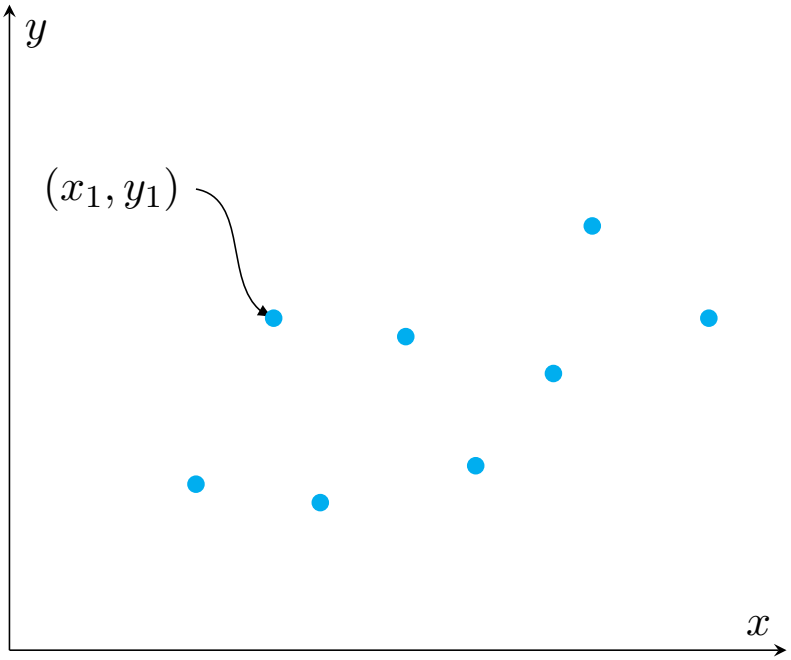


Fig 1: Given Sample Set

We will find a line that passes through these points, there by being the best line, that has minimum vertical or Δy distance from all the sample points. Typically such a line would be unique to given any sample set and it is the **best fit** line possible. Figure 2.2 shows such a *potential* line. The vertical difference Δy_1 as shown in figure, is the distance between the point (x_1, y_1) and the line.

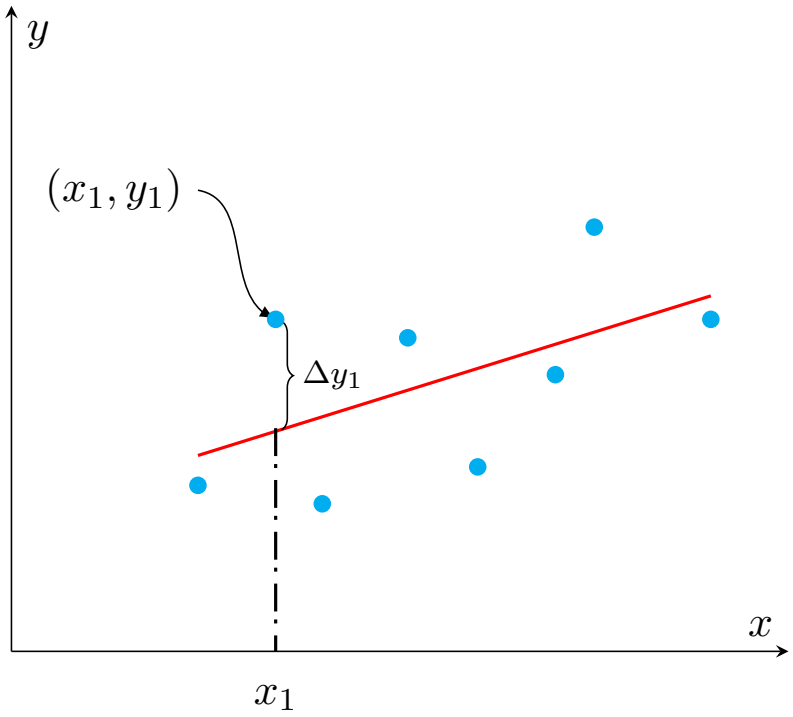


Fig 2: Finding a best-fit line is the goal

When a sample set is given, we will assume such a line exists and that ideally, all sample points should have fallen on that line, implying a perfect linear relationship between x and y. However, because of an **underlying error** ε, the sample points have fallen apart, around the line, giving us the sample set. Suppose, such a perfect linear relationship exists ideally, let us say, it could be defined as below by using a regular line equation with slope β_1 and y-intercept β_0, as

$$y = \beta_0 + \beta_1 x$$

Thus in this ideal world, y is completely deterministic from x. However, when we introduce randomness in the form or error ε, the y value also becomes a random variable associated with the randomness from ε. That is, if we describe such a RV as Y, then

$$Y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

We do not know ε. Naturally, the *expectation* of the error is to be zero, or in other words we assume, though there is room for error, but the zero error has maximum probability. Thus assuming a normal distribution of $N(0, \sigma^2)$,

$$E(\varepsilon) = 0 \quad Var(\varepsilon) = \sigma^2 \tag{2.2}$$

Assumptions and Approach

- Given the sample set value, we will imagine there to be an *ideal* linear relationship, and try to find that hypothetical line which will have *least error* for all observed sample points.
- Also assume that error has maximum probability to be 0, and normally distributed, formulating as a cause for observing sample values as observed instead of, on the line where error would have been zero.

This line of thought is important and fundamental to our model. Because of this assumption, we could now say, the points should have ideally sat on the line, but resulted in their places in reality as we find them, because of the error. Thus the observed y value is the result of the error ε, while its **expected y value** $E(Y|x)$ **or** $\mu_{Y.x_1}$, should sit on the line. This is illustrated in Figure 2.3. Similarly the only randomness comes from error ε, so its variance directly transfers to the Y random variable due to 2.2. That is, $\sigma_{Y.x_1} \to \sigma$.

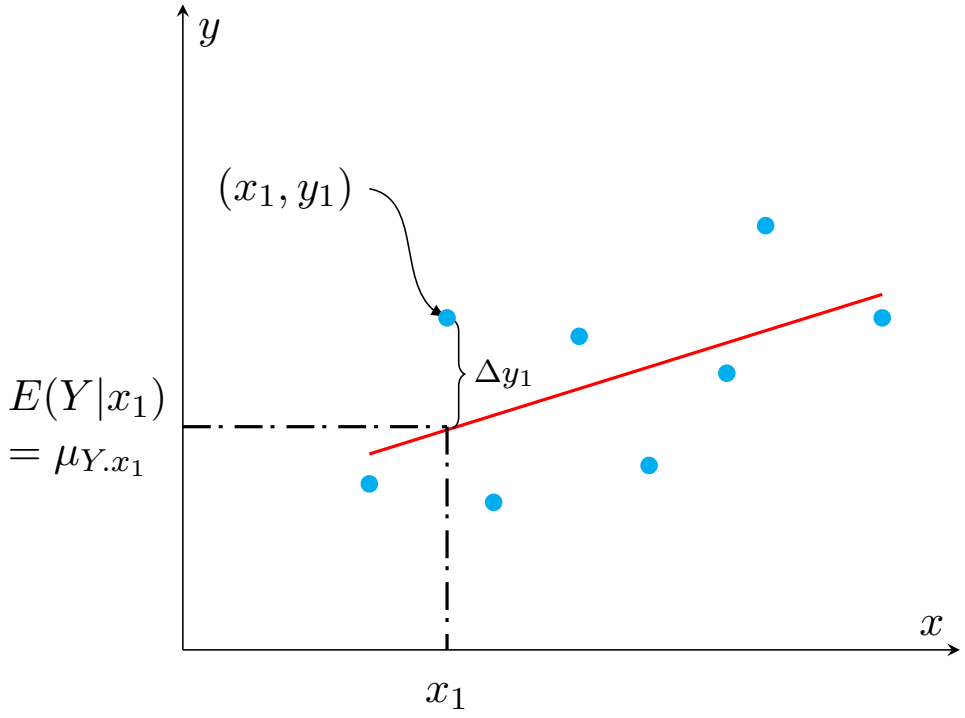


Fig 3: $y1$ and $E(Y|x_1)$

We could also prove them mathematically as below. For any point (x_1, y_1)

$$\mu_{Y.x_1} = E(Y|x_1) = E(\beta_0 + \beta_1 x_1 + \varepsilon) = E(\beta_0) + E(\beta_1 x_1) + E(\varepsilon)$$

If a is a constant observed, then $E(a) = a$ only as its the only value and already observed. And since $\varepsilon = N(0, \sigma^2)$ we could write,

$$\mu_{Y.x_1} = E(Y|x_1) = \beta_0 + \beta_1 x_1$$

Similary,

$$\sigma^2_{Y.x_1} = Var(Y|x_1) = Var(\beta_0 + \beta_1 x_1 + \varepsilon)$$

If a is a constant observed, then $Var(a) = 0$ only as its already observed and there is no uncertainty. And since $\varepsilon = N(0, \sigma^2)$ we could write,

$$\sigma^2_{Y.x_1} = Var(Y|x_1) = 0 + 0 + \sigma^2$$

Thus, in general for any x, in continuous scale, we could say,

$$\mu_{Y.x} = E(Y|x) = \beta_0 + \beta_1 x$$
$$\sigma^2_{Y.x} = Var(Y|x) = \sigma^2$$

Note, though our sample values are discrete, we are able to get a line at continuous scale, because its the ideal situation, where all the expected values should lie on that hypothetical line $y = \beta_0 + \beta_1 x$. So this line should stay true for any value of x. It is a hypothetical line of expected or mean values $E(Y|x)$, so understandably, its called **line of mean values**. It should also have been the ideal line, where all sample points should have rested, provided there were no errors. So this line is also called **True regression line**.

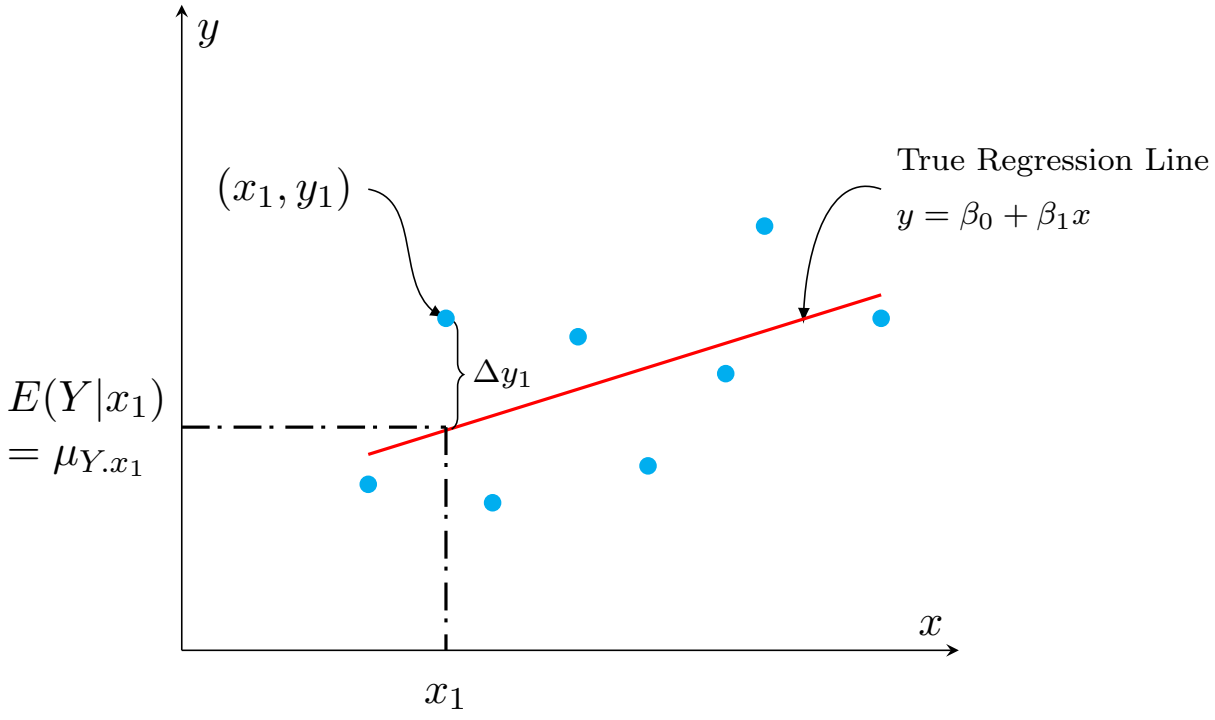


Fig 4: $\beta_0 + \beta_1 x$ is the ideal hypothetical line with no error

Expected value and Variance of Y given a sample x^*

For any observed value (x^*, y^*),

$$\begin{aligned} \mu_{Y.x^*} &= E(Y|x^*) = \beta_0 + \beta_1 x^* \\ \sigma^2_{Y.x^*} &= Var(Y|x^*) = \sigma^2 \end{aligned} \tag{2.3}$$

In continuous scale, for any (x, y),

$$\begin{aligned} \mu_{Y.x} &= E(Y|x) = \beta_0 + \beta_1 x \\ \sigma^2_{Y.x} &= Var(Y|x) = \sigma^2 \end{aligned} \tag{2.4}$$

It is difficult to visualize the error randomness (say, its pdf) in the x, y graph as ε is another 3rd variable hidden underneath. However we just saw, how that distribution transfers to the random variable Y. If ε has $N(0, \sigma^2)$, then Y has distribution $N(\beta_0 + \beta_1 x, \sigma^2)$. This facilitates us to view the randomness on the face of random variable Y as shown in 2.5. Observe that, for a point, say (x_1, y_1), for the given x_1, ideally, y should have been the mean value $E(Y|x_1) = \beta_0 + \beta_1 x$, that has the highest probability of the normal distribution. That is our assumption and then we say, because there exists an error, we got y at y_1. Note for the sample location y_1, the error is low, but still had a chance.

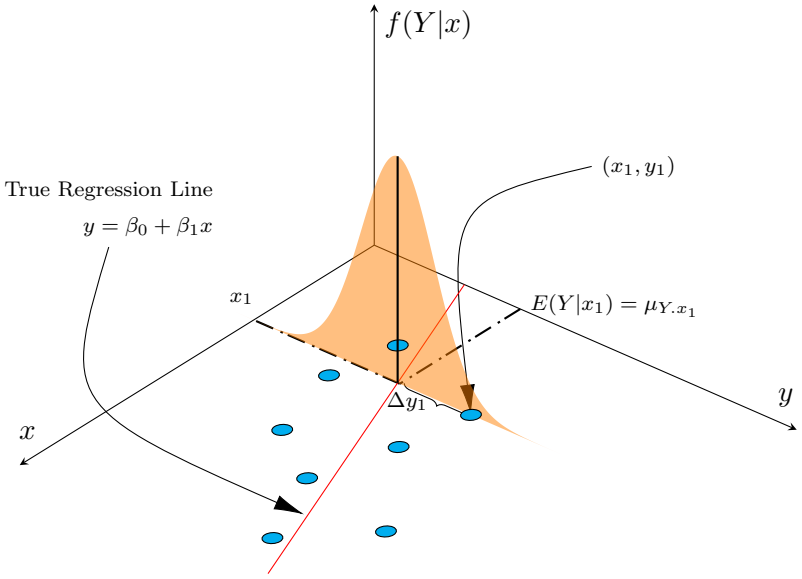


Fig 5: The Probability Distribution $f(Y|x_1)$

The distance how much the *erroneous* locations of sample points spread out from the mean value is determined by variance σ of the error. Note that, we assume this error is constant for all sample values. This means, any point x_m, y_m has same probability distribution of committing an error, as any other point in the sample set. This assumed property is called **Homoscedasticity**. If this is not the case, then the characteristic is called **Heteroscedasticity**. One could fairly assume from given a sample set, if the underlying error could be Homoscedastic or Heteroscedastic, by eyeballing at the spread from the regression line. We will focus and assume Homoscedasticity and for any one interested, Frost [2] has written an interesting article about dealing with the same. Given that Homoscedasticity is assumed, the probability distribution would be uniform across the

regression line. This is illustrated in 2.6. That is, for any x value, the equivalent $f(Y|x)$ could be picked up like a card from a stack. This distribution across the regression line could be continuous or discrete, depending on x is continuous or discrete.

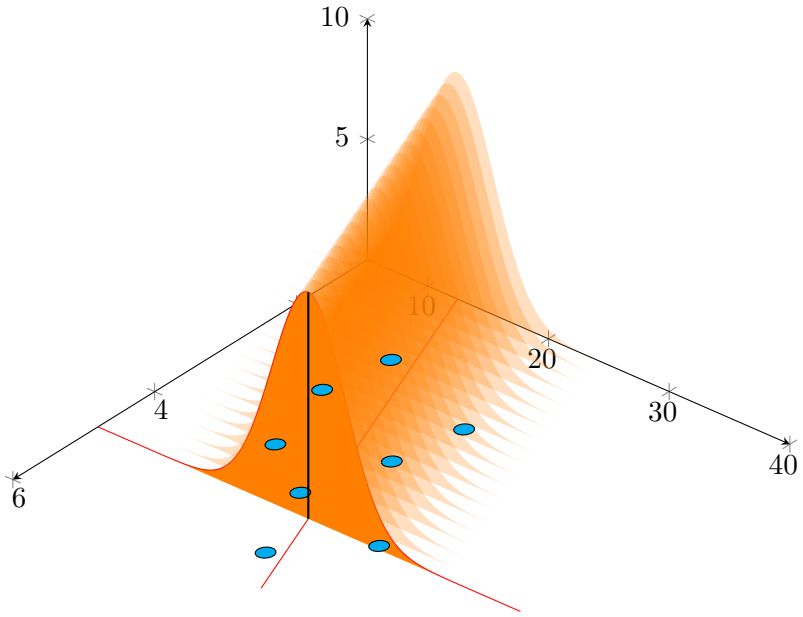


Fig 6: The *pdf* $f(Y|x)$ is continuous or discrete along the regression line depending on x is continuous or discrete

Now that our sample set is discrete, let us focus on that. We need to find out, for given sample set, what would be the optimal values of β_0 and β_1.

2.1.2 Estimating Model Parameters

The goal is to find (β_0, β_1) such that, the resulting line is some how "best-fit" among all possible lines of $E(Y|x)$. You see, our sample set could be a part of a bigger population, and thus the hypothetical line for entire population could be anything. However, we have only a sample set, so our best bet is always what is the best representative of the sample. That is, **given the samples**, what would be the best representative regression line is what our goal is. Imagine, if all sample lines, line up in a certain way, then our best bet would be just a line cutting across all those points. This suggests, all sample points have zero error, or have fallen at their respective highest probability mean locations, thus one could expect any more new sample to take a similar place on that line. Note that in this case, all lines are at *zeroth distance* from the mean line. This is illustrated in 2.7 where the vertical red dotted line represents maximum probability.

Now when the samples deviate from such a hypothetical mean line, best bet then to find the mean line is to find one, that has *least distance* from all the sample points. The sum of all the distances from all sample points to that line would be minimal compared to any other lines' similar sum of distances. The distances are illustrated in 2.8, where blue lines indicate the actual distance from the true regression line. Now, naturally, since the points could lie on either side of the line, would give rise to relatively positive or negative distances, and thus cancelling each others' distances out partly here and there. To avoid that, one could take absolute distances from the point to the line.

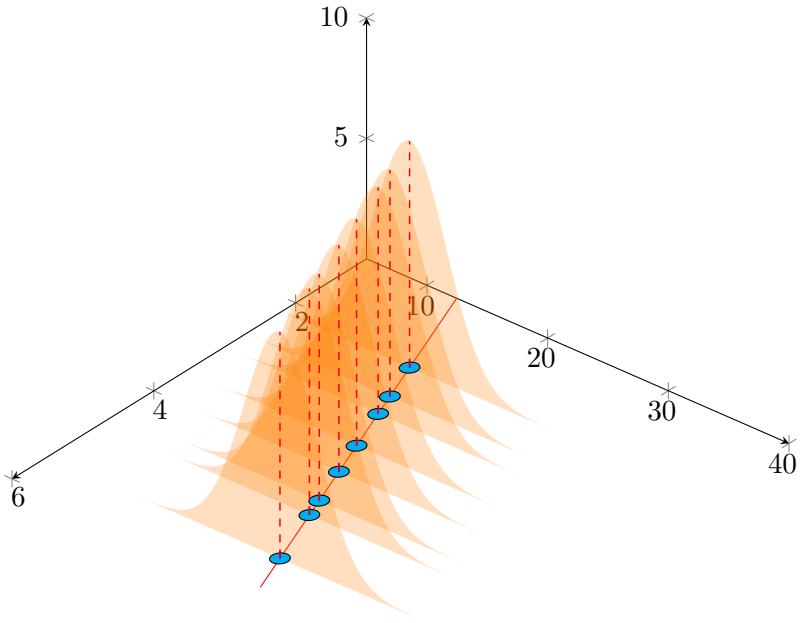


Fig 7: An ideal case

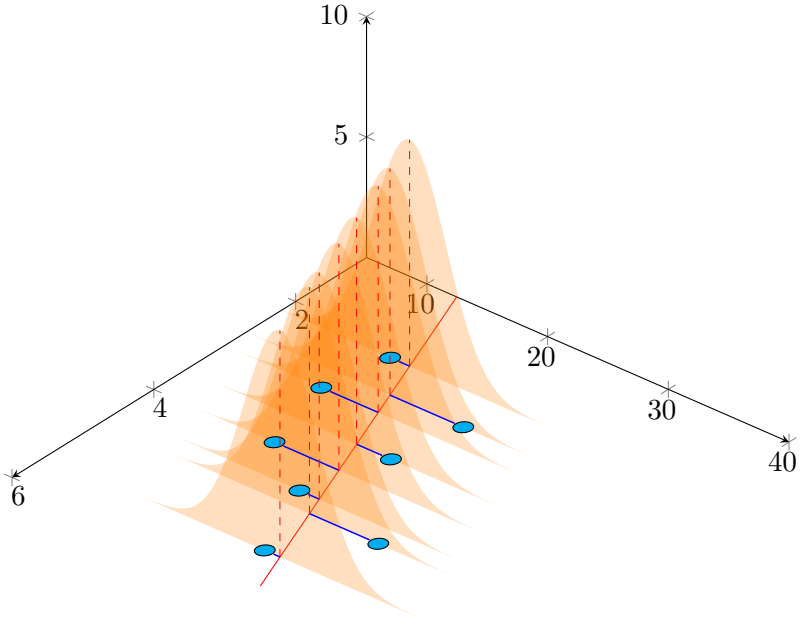


Fig 8: A practical case

Principle of Least Squares

However, instead of taking the absolute distances, we now, out of nowhere(?!) choose to take the square of the calculated distance and sum up to find the total distance. As per my current understanding, this was mearly a choice for algebraic convenience [1]. We also have other ways of measuring approaches (angled distance instead of vertical etc) but we shall not get in to it as this is only Simple Regression Model.

Now that we have fixated on finding the least sum of squares of the distances (note because we squared, there was no absoluteness to be considered in equation), let us look in to the mathematical form of it. This principle which can be traced back to famous mathematician Guass, says that, a line provides a good fit to the data if the vertical distances (deviations) from the observed points to the line are small. The measure of the goodness of fit is the sum of the squares of these deviations. The best-fit line is then the one having the smallest possible sum of squared deviations.

Principle of Least Squares (from Devore [ ])

The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1 x$ is

$$\text{height of point - height of line} = y_i - (b_0 + b_1 x_i)$$

The sum of squared vertical deviations from the points $(x_1, y_1), \cdots, (x_m, y_m)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2 \tag{2.5}$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least square estimates**, are those values that minimize $f(b_0, b_1)$. That is $(\hat{\beta}_0, \hat{\beta}_1)$ are such that, $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ for any $(b_0.b_1)$. The **estimated regression line or least squares line** is then the line whose equation is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.6}$$

Note that 2.6 is same as expected mean line or true regression line as expressed in 2.4. Here we just devised a way to find those optimal (β_0, β_1).

Using Maximum Likelihood Estimation

We could also arrive at 2.5 via Maximum Likelihood Estimation (which was the reason we had entire chapter on MLE before regression in first place). Recall each sample point as shown on figure 2.8, has the pdf $f(Y|x) = N(\beta_0 + \beta_1 x, \sigma^2)$. Then, as per MLE, we would like to know what is the joint probability of all these samples points to be at their observed locations. It will be useful to recall MLE derivation for Normal distribution as we saw in 1.12. In similar fashion, for each sample point, the pdf could be written as,

$$f(Y|x_i; \beta_0, \beta_1) = N(\beta_0 + \beta_1 x_i, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ - \frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \Big\}$$

And as usual, assuming all these sample points are *independent and identically distributed*, we could arrive at their likelihood function as

[1] http://www.bradthiessen.com/html5/docs/ols.pdf

$$
\begin{aligned}
L(\beta_0, \beta_1) &= f(Y|x_1; \beta_0, \beta_1, Y|x_2; \beta_0, \beta_1, \cdots Y|x_m; \beta_0, \beta_1) \\
&= f(Y|x_1; \beta_0, \beta_1) f(Y|x_2; \beta_0, \beta_1) \cdots f(Y|x_m; \beta_0, \beta_1) \\
&= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ - \frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \Big\} \\
&= \left\{ \Big(\frac{1}{2\pi\sigma^2} \Big)^{\frac{m}{2}} \right\} \left\{ \prod_{i=1}^{m} \exp\Big\{ - \frac{[y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \Big\} \right\} \\
&= \big(2\pi\sigma^2\big)^{\frac{-m}{2}} \left\{ \exp\Big\{ - \frac{\sum_{i=1}^{m} [y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \Big\} \right\}
\end{aligned}
$$

Note, using product rule of logarithms, for any function $f = p^a e^b$,

$$ln(p^a e^b) = a ln(p) + b$$

Thus, taking natural logarithm on both sides of likelihood function,

$$ln(L(\beta_0, \beta_1)) = -\frac{m}{2}\big(ln(2\pi\sigma^2)\big) - \frac{\sum_{i=1}^{m} [y_i - (\beta_0 + \beta_1 x_i)]^2}{2\sigma^2} \tag{2.7}$$

The function 2.7 is a function of two variables (β_0, β_1), thus graphically represents a 3D surface plot as shown in figure 2.9, with height of the surface at any point is the function value evaluated at that point. We need to find out a point on this surface, where the function reaches maximum. The value of (β_0, β_1) at that point represents optimal values $(\hat{\beta}_0, \hat{\beta}_1)$. Why? Because, associated with those points, is the probability density function that yields maximum probability of getting all those sample sets in the places they are observed.

MLE leads to OLS

Before we find the optimal points, note that equation 2.7 has the variables (β_0, β_1) in the second term of RHS, and thus it is on that term we would be operating upon to find the optimal value. That is, when we derive w.r.t. (β_0, β_1) , the first term on RHS is a constant so goes away and constants in 2nd term too, would not offer any information, which we will see shortly, due to which we would just be equating the numerator of 2nd term RHS, to find the optimal value. That is, let

$$H(\beta_0, \beta_1) = \sum_{i=1}^{m} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{2.8}$$

then, by attempting to find the critical points of log likelihood $lnL(\beta_0, \beta_1)$ of given sample set, we would essentially operate upon $H(\beta_0, \beta_1)$. Note that this $H(\beta_0, \beta_1)$ is exactly equivalent to the ordinary least squares equation we saw in 2.5.

Derivation

To find the critical points on the surface (which could be maximum or minimum or saddle point), let us take first order partial derivatives and equate to 0. For details on why we do this, refer appendix 3.2.7 where we have shortly explained the concept behind using derivatives for finding critical points.

Keeping β_0 as constant and taking partial derivative with respect to β_1, we get,

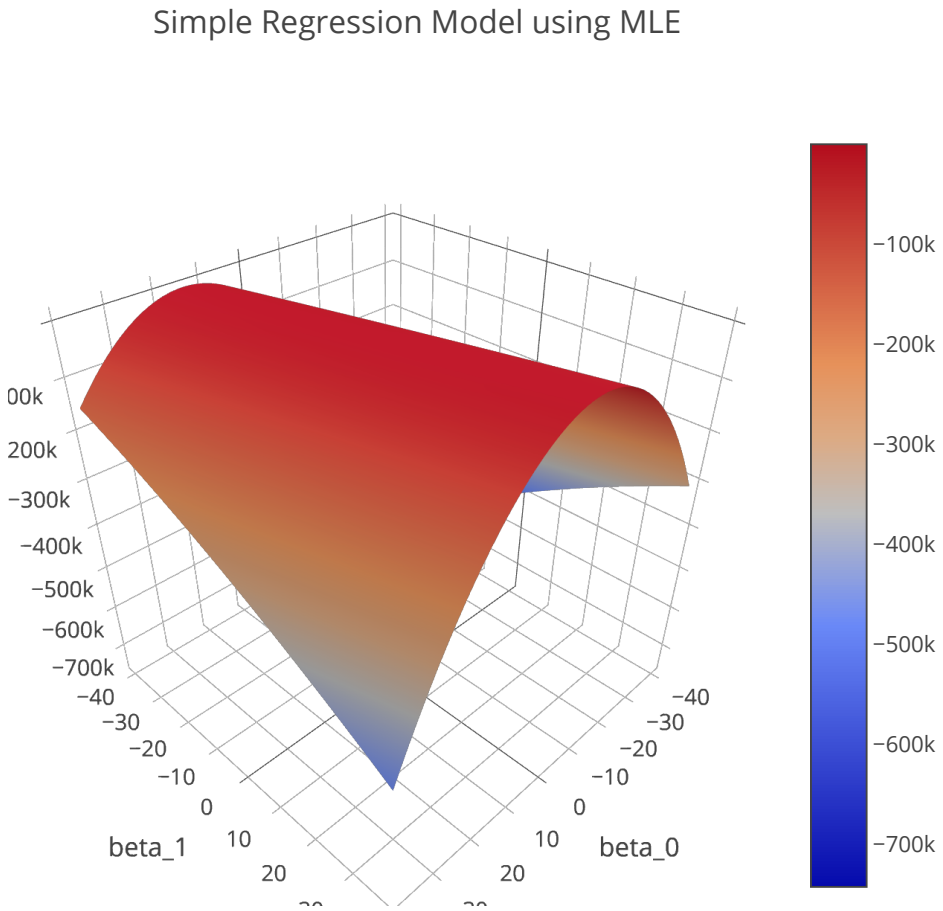


Fig 9: Log Likelihood function of given sample set

$$\left.\frac{\partial ln(\beta_0,\beta_1)}{\partial \beta_1}\right|_{\beta_0=k} = 0 - 2\Big\{\frac{\sum_{i=1}^{m}[y_i-(\beta_0+\beta_1 x_i)(-x_i)]}{2\sigma^2}\Big\}$$
$$= \frac{1}{\sigma^2}\Big\{\sum_{i=1}^{m}[y_i-\beta_0-\beta_1 x_i](x_i)\Big\}$$

Keeping β_1 as constant and taking partial derivative with respect to β_0, we get,

$$\left.\frac{\partial ln(\beta_0,\beta_1)}{\partial \beta_0}\right|_{\beta_1=k} = 0 - 2\Big\{\frac{\sum_{i=1}^{m}[y_i-(\beta_0+\beta_1 x_i)]}{2\sigma^2}\Big\}(-1)$$
$$= \frac{1}{\sigma^2}\Big\{\sum_{i=1}^{m}[y_i-\beta_0-\beta_1 x_i]\Big\}$$

Equating both to 0, we get, (note, now the paramters are $(\hat{\beta}_0,\hat{\beta}_1)$) because they are the optimal *values* we are going to find out by equating to 0.

$$\sum_{i=1}^{m}[y_i-\hat{\beta}_0-\hat{\beta}_1 x_i] = 0 \tag{2.9}$$
$$\sum_{i=1}^{m}[y_i-\hat{\beta}_0-\hat{\beta}_1 x_i]x_i = 0 \tag{2.10}$$

Due to repeated use, for a while, let $\sum_{i=1}^{m} \implies \sum_i$.

We know $\overline{x} = \frac{1}{m}\sum_i x_i$, and $\overline{y} = \frac{1}{m}\sum_i y_i$. Thus,

$$\sum_i x_i = m\overline{x} \tag{2.11}$$
$$\sum_i y_i = m\overline{y} \tag{2.12}$$

Substituting in 2.9,

$$\sum_i[y_i-\hat{\beta}_0-\hat{\beta}_1 x_i] = 0$$
$$\sum_i y_i - m\hat{\beta}_0 - \hat{\beta}_1\sum_i x_i = 0$$
$$m\overline{y} - m\hat{\beta}_0 - m\hat{\beta}_1\overline{x} = 0$$
$$\overline{y} - \hat{\beta}_0 - \hat{\beta}_1\overline{x} = 0$$
$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1\overline{x} \tag{2.13}$$

For any x_i, let

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2.14}$$

Substituting 2.14 in 2.10,

$$\sum_i [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] x_i = 0$$
$$\sum_i [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] x_i = 0$$
$$\sum_i (y_i - \hat{y}_i) x_i = 0 \tag{2.15}$$

Solving for β_1

Subtract 2.13 from 2.14,

$$\begin{aligned} \hat{y}_i - \overline{y} &= (\hat{\beta}_0 + \hat{\beta}_1 x_i) - \hat{\beta}_0 + \hat{\beta}_1 \overline{x} \\ &= \hat{\beta}_1 (x_i - \overline{x}) \end{aligned} \tag{2.16}$$

Adding and cancelling y_i on LHS,

$$(\hat{y}_i - \overline{y}) + (y_i - y_i) = \hat{\beta}_1 (x_i - \overline{x})$$
$$(\hat{y}_i - y_i) + (y_i - \overline{y}) = \hat{\beta}_1 (x_i - \overline{x})$$

Multipying both sides by $(x_i - \overline{x})$ and summing up

$$(\hat{y}_i - y_i)(x_i - \overline{x}) + (y_i - \overline{y})(x_i - \overline{x}) = \hat{\beta}_1 (x_i - \overline{x})(x_i - \overline{x})$$
$$\sum_i (\hat{y}_i - y_i)(x_i - \overline{x}) + \sum_i (y_i - \overline{y})(x_i - \overline{x}) = \hat{\beta}_1 \sum_i (x_i - \overline{x})^2 \tag{2.17}$$

Focussing on $\sum_i (\hat{y}_i - y_i)(x_i - \overline{x})$

$$\sum_i (\hat{y}_i - y_i)(x_i - \overline{x}) = \sum_i (\hat{y}_i - y_i) x_i - \overline{x} \sum_i (\hat{y}_i - y_i)$$

Note from 2.15, $\sum_i (\hat{y}_i - y_i) x_i$ is 0. Thus,

$$\sum_i (\hat{y}_i - y_i)(x_i - \overline{x}) = -\overline{x} \sum_i (\hat{y}_i - y_i)$$

Let us calculate $\sum_i (\hat{y}_i - y_i)$ separately,..

$$\begin{aligned}\sum_i(\hat{y}_i - y_i) &= \sum_i \hat{y}_i - \sum_i y_i \\ &= \sum_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) - m\overline{y} \\ &= \sum_i \hat{\beta}_0 + \sum_i \hat{\beta}_1 x_i - m\overline{y} \\ &= m\hat{\beta}_0 + m\hat{\beta}_1\overline{x} - m\overline{y} \\ &= m(\hat{\beta}_0 + \hat{\beta}_1\overline{x}) - m\overline{y} \\ &= m\overline{y} - m\overline{y} \\ &= 0 \end{aligned} \tag{2.18}$$

Thus,

$$\sum_i(\hat{y}_i - y_i)(x_i - \overline{x}) = 0 \tag{2.19}$$

Substituting 2.19 in 2.17,

$$\begin{aligned}\sum_i(y_i - \overline{y})(x_i - \overline{x}) &= \hat{\beta}_1 \sum_i(x_i - \overline{x})^2 \\ \implies \hat{\beta}_1 &= \frac{\sum_i(y_i - \overline{y})(x_i - \overline{x})}{\sum_i(x_i - \overline{x})^2}\end{aligned}$$

From 2.13,

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$$

Regression Parameters using MLE

For the true line of regression $E(Y|x) = \hat{\beta}_0 + \hat{\beta}_1 x$,

$$\hat{\beta}_1 = \frac{\sum_i(y_i - \overline{y})(x_i - \overline{x})}{\sum_i(x_i - \overline{x})^2} \tag{2.20}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x} \tag{2.21}$$

It is strongly advised to check out our interactive example [2] where we have shown visually and also proven how close the results are, between direct formula we just derived and also if directly picking up point of maximum value from the log likelihood graph itself.

[2] http://nbviewer.jupyter.org/gist/parthi2929/e092970b94ee6aeb99519457df41921a

Chapter 3

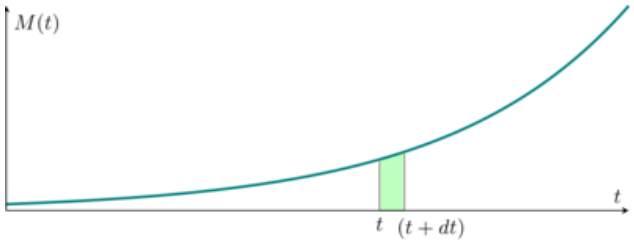
Appendix

3.1 e and natural logarithms

3.1.1 The basics of e

Case 1:

Suppose we have a function $M(t) = 2^t$ and we are interested in its rate of change $\frac{dM(t)}{dt}$.



$$\frac{dM(t)}{dt} = \lim_{dt\to 0} \frac{2^{(t+dt)} - 2^t}{dt} = \lim_{dt\to 0} \frac{2^t 2^{dt} - 2^t}{dt} = \lim_{dt\to 0} \frac{2^t(2^{dt} - 1)}{dt} \therefore \frac{d(2^t)}{dt} = \lim_{dt\to 0} 2^t\Big(\frac{2^{dt} - 1}{dt}\Big)$$

One could note that, as $dt \to 0$, the component $\Big(\frac{2^{dt} - 1}{dt}\Big) \to 0.6931$ as shown below.

```
In[36]: dt = [0.1, 0.01, 0.005, 0.001,0.0005, 0.0001]
        cl = [print(i, round((2**i-1)/i,5)) for i in dt]
```

```
0.1 0.71773
0.01 0.69556
0.005 0.69435
0.001 0.69339
0.0005 0.69327
0.0001 0.69317
```

So,

$$\frac{d(2^t)}{dt} = \lim_{dt\to 0} 2^t\Big(\frac{2^{dt} - 1}{dt}\Big) = 2^t(0.6931) \tag{3.1}$$

Case 2:

Suppose we have a function $M(t) = 3^t$ and we are interested in its rate of change $\frac{dM(t)}{dt}$.

$$\frac{dM(t)}{dt} = \lim_{dt\to 0} \frac{3^{(t+dt)} - 3^t}{dt} = \lim_{dt\to 0} \frac{3^t 2^{dt} - 3^t}{dt} = \lim_{dt\to 0} \frac{3^t(2^{dt} - 1)}{dt} \therefore \frac{d(3^t)}{dt} = \lim_{dt\to 0} 3^t\Big(\frac{3^{dt} - 1}{dt}\Big)$$

One could note that, as $dt \to 0$, the component $\Big(\frac{3^{dt} - 1}{dt}\Big) \to 1.09867$ as shown below.

```
In[38]: dt = [0.1, 0.01, 0.005, 0.001,0.0005, 0.0001]
        cl = [print(i, round((3**i-1)/i,5)) for i in dt]
```

```
0.1 1.16123
0.01 1.10467
0.005 1.10164
0.001 1.09922
0.0005 1.09891
0.0001 1.09867
```

So,

$$\frac{d(3^t)}{dt} = \lim_{dt\to 0} 3^t\Big(\frac{3^{dt} - 1}{dt}\Big) = 3^t(1.09867) \tag{3.2}$$

Generalization

Similarly for any $M(t) = a^t$, we could prove,

$$\frac{d(a^t)}{dt} = \lim_{dt\to 0} a^t\Big(\frac{a^{dt} - 1}{dt}\Big) = a^t C \quad \text{where C is some constant} \tag{3.3}$$

Wonder

Naturally if we wonder, is there any similar $M(t)$ for which the derivative is itself? (In other words, that some constant becomes 1?!). We could solve this as below.

We want to find a such that,

$$\lim_{dt\to 0} \Big(\frac{a^{dt} - 1}{dt}\Big) = 1$$

Rewriting,

$$\lim_{dt\to 0} a^{dt} = 1 + dt \therefore a = \lim_{dt\to 0} (1 + dt)^{1/dt}$$

We can mathematically prove that, $(1 + n)^{1/n}$ approaches a constant, but for here, we could simply compute like earlier, what is the value it is approaching..

```
In[45]: dt = [0.1, 0.01, 0.005, 0.001,0.0005, 0.0001]
        cl = [print(i, round((1 + i)**(1/i),5)) for i in dt]
```

```
0.1 2.59374
0.01 2.70481
0.005 2.71152
0.001 2.71692
0.0005 2.7176
0.0001 2.71815
```

$\therefore$ we do have one constant 2.718 for which, the derivative of it is itself. That is,

The value of e

Let $e = 2.71815$, then

$$\frac{d(e^t)}{dt} = e^t \tag{3.4}$$

3.1.2 Derivative of e^{ct}

What is $\frac{d(e^{ct})}{dt}$? This can be solved by chain rule in differential calculus.
Let $u = ct$, then by chain rule,

$$\frac{d(e^{ct})}{dt} = \frac{d(e^u)}{dt} = \frac{d(e^u)}{du}\frac{du}{dt} = e^u\frac{du}{dt}$$

Substituting $u = ct$ back,

$$\frac{d(e^{ct})}{dt} = e^{ct}\frac{d(ct)}{dt} = ce^{ct}$$

The derivative of e^{ct}

Let $e = 2.71815$, then

$$\frac{d(e^{ct})}{dt} = ce^{ct} \tag{3.5}$$

3.1.3 Using e for any exponent form

A short summary of what we saw earlier.

$$\frac{d(2^t)}{dt} = (0.6931)2^t$$
$$\frac{d(3^t)}{dt} = (1.0986)3^t$$
$$\frac{d(a^t)}{dt} = (C)a^t, \quad \text{where } C \text{ is some constant depending on } a$$

Let $2 = e^C$. Then

$$2^t = e^{Ct}$$

Taking derivatives on both sides,

$$\frac{d(2^t)}{dt} = \frac{d(e^{Ct})}{dt} \implies (0.6931)2^t = Ce^{Ct} \implies C = 0.6931$$

That is, the constant we earlier got, is nothing but the power to which we need to raise e to get the base value 2. That is, $2 = e^{0.6931}$. We could call this constant as **natural logarithm of 2**, denoted by $ln(2)$ or $log_e(2)$

Similarly, let $3 = e^C$. Then

$$3^t = e^{Ct}$$

Taking derivatives on both sides,

$$\frac{d(3^t)}{dt} = \frac{d(e^{Ct})}{dt} \implies (1.0986)2^t = Ce^{Ct} \implies C = 1.0986$$

Thus, $3 = e^{1.0986}$. We could call this constant as **natural logarithm of 3**, denoted by $ln(3)$ or $log_e(3)$

Summarizing,

$$2 = e^{ln(2)}, \quad ln(2) = log_e(2) = 0.6931$$
$$3 = e^{ln(3)}, \quad ln(3) = log_e(3) = 1.0986$$

Any number in terms of e

Any number could be equated by e to the power of its natural logarithmic value, which is a unique constant that could be derived.

$$a = e^{ln(a)}, \quad ln(a) = log_e(a) \tag{3.6}$$

3.1.4 Multiplication and Division simplified

Suppose we have a function $L(p, q) = p^y q^z$

We could make the multiplication of such exponents in to simpler form of addition of their natural logarithms as below.

Let $p = e^{C_1}$ and $q = e^{C_2}$, then we already have seen, $C_1 = ln(p), C_2 = ln(q)$.

$$\therefore p^y q^z = e^{C_1 y} e^{C_2} z = e^{C_1 y + C_2 z} = e^{ln(p)y + ln(q)z}$$

If $L = e^{ln(L)}$ similarly, then we could write,

$$\begin{aligned} L &= p^y q^z \\ e^{ln(L)} &= e^{ln(p)y + ln(q)z} \\ \implies ln(L) &= yln(p) + zln(q) \quad \text{or} \\ log_e(L) &= ylog_e(p) + zlog_e(q) \end{aligned}$$

If $L(p, q) = \dfrac{p^y}{q^z}$

$$\frac{p^y}{q^z} = \frac{e^{C_1 y}}{e^{C_2 z}} = e^{C_1 y - C_2 z} = e^{ln(p)y - ln(q)z}$$

If $L = e^{ln(L)}$ similarly, then we could write,

$$L = \frac{p^y}{q^z}$$
$$e^{ln(L)} = e^{ln(p)y - ln(q)z}$$
$$\implies ln(L) = yln(p) - zln(q) \quad \text{or}$$
$$log_e(L) = ylog_e(p) - zlog_e(q)$$

Thus we have simplified multiplication and division to addition and subtraction provided we know the equivalent natural logarithms of the values involved.

Multiplication and Divison Simplification

- If $L(p, q) = p^y q^z$, then
$$log_e(L) = ylog_e(p) + zlog_e(q) \tag{3.7}$$
- If $L(p, q) = \frac{p^y}{q^z}$, then
$$log_e(L) = ylog_e(p) - zlog_e(q) \tag{3.8}$$

3.1.5 Derivatives of ln

We only see few derivatives that could be useful in MLE.

Q1: What is the derivative of $\frac{d(ln(x))}{dx}$?
Let $y = ln(x) = log_e x$. This means, $e^y = x$
Differentiating that,

$$e^y = x\frac{d(e^y)}{dx} = \frac{dx}{dx}e^y\frac{dy}{dx} = 1\frac{dy}{dx} = \frac{1}{e^y} \therefore \frac{d(ln(x))}{dx} = \frac{1}{x}$$

Q2: What is the derivative of $\frac{d(ln(1-x))}{dx}$?
Let $y = ln(1-x) = log_e(1-x)$. This means, $e^y = 1 - x$
Differentiating that,

$$e^y = 1 - x\frac{d(e^y)}{dx} = \frac{d(1-x)}{dx}e^y\frac{dy}{dx} = -1\frac{dy}{dx} = \frac{-1}{e^y} = \frac{-1}{1-x} \therefore \frac{d(ln(1-x))}{dx} = \frac{-1}{1-x}$$

Q3: What is the derivative of $\frac{d(ln(cx))}{dx}$?
Let $u = cx$, $y = ln(u)$. This means, $e^y = u$
Differentiating that,

$$e^y = u$$
$$\frac{d(e^y)}{dx} = \frac{du}{dx} = \frac{d(2\pi x)}{dx} = 2\pi$$
$$e^y\frac{dy}{dx} = 2\pi$$
$$\frac{dy}{dx} = \frac{2\pi}{e^y} = \frac{2\pi}{u} = \frac{2\pi}{2\pi x} = \frac{1}{x}$$

Derivatives of ln

- $$\frac{d(ln(x))}{dx} = \frac{1}{x} \tag{3.9}$$
- $$\frac{d(ln(1-x))}{dx} = \frac{-1}{1-x} \tag{3.10}$$
- $$\frac{d(ln(cx))}{dx} = \frac{1}{x} \tag{3.11}$$

3.2 Applying derivatives to analyze functions

3.2.1 Introduction

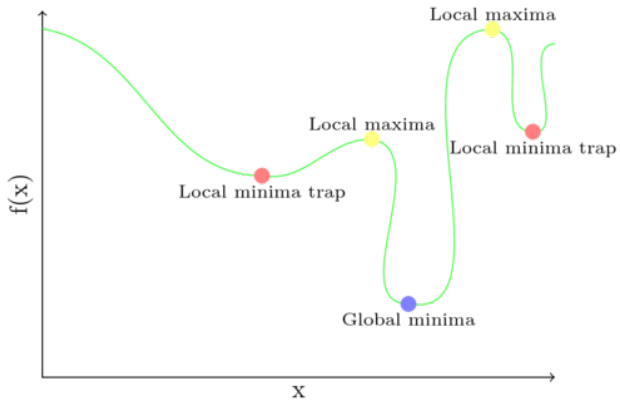
Inspired from this video from Khan Academy.

Here we will have a quick glance on the basics of finding maxima and minima of a given function.

Maxima - plural of maximum Minima - plural of minimum $[a, b]$ - closed interval - includes a and b (a, b) - open interval - excludes a and b

3.2.2 Critical Points

Given a function $f(x)$,



1. There could be a **global maximum and a minimum** point. Global in the sense, inside the *interval*, it is the maximum or minimum out of all peaks or valleys.
2. There are areas, which are **local maximum or minium** around those areas. There could more than one local maximum or minimum points within the interval.
3. The slopes at these points are 0 or also *undefined* (sharp turn), that is $f'(x) = 0$ at these points. All these points are called **critical points**.
4. Suppose the interval is [a,b]. Then the **end points** are a and b are not critical points, because anyway $f'(a)$ and $f'(b)$ would be 0 or undefined.

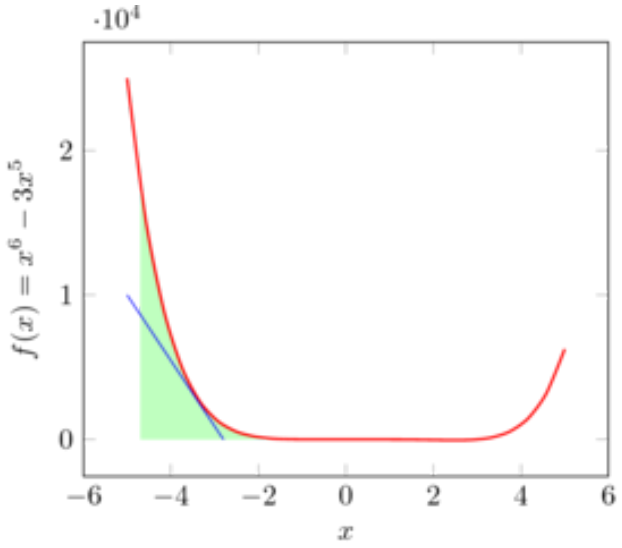
5. Not all critcal points, which have slope 0, becomes a global or local maximum or minimum point. It could have just flattened.

3.2.3 Decreasing or Increasing Interval

Inspired from this video from Khan Academy.

Decreasing interval

Suppose $f(x) = x^5(x-3)$. Below is how the function looks like. We need to find the interval within which the function is decreasing. Just by eyeballing, we could know that the green area is where the function is decreasing, but we are not clear of the exact intervals. This would could find mathematically.

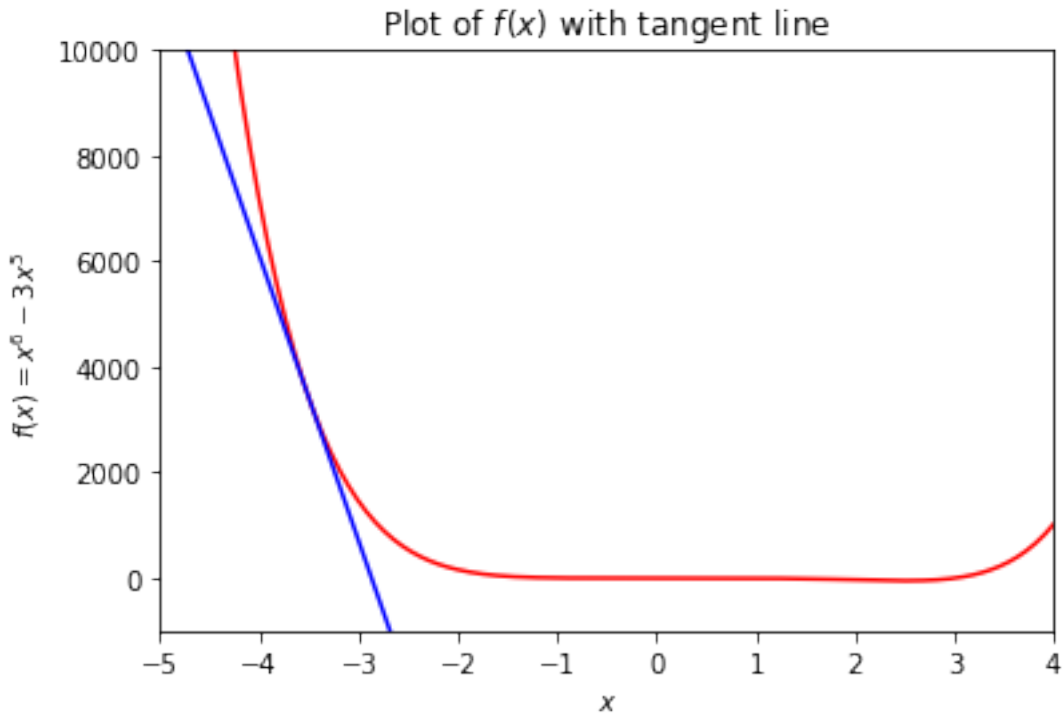


Note that, any tangent line drawn on the curve in the green area will have a negative slope as shown by a blue line. This means, $f'(x) < 0$ in those areas. This is the clue. Finding the derivatives, we get

$$f'(x) = 6x^5 - 15x^4$$

$$f'(x) < 0 \implies (6x^5 - 15x^4) < 0 \implies (3x^4)(2x-5) < 0$$

But $3x^4 > 0$ always for any x due to even power, so only possibility should be $(2x-5) < 0$ or $x < 2.5$. This means the exact interval where function is decreasing is $-\infty < x < 2.5$. Below is the python implementation of the curve with tangent.



Similarly we could also see where $x > 2.5$, the curve is increasing. Summarizing,

Decreasing or Increasing Interval

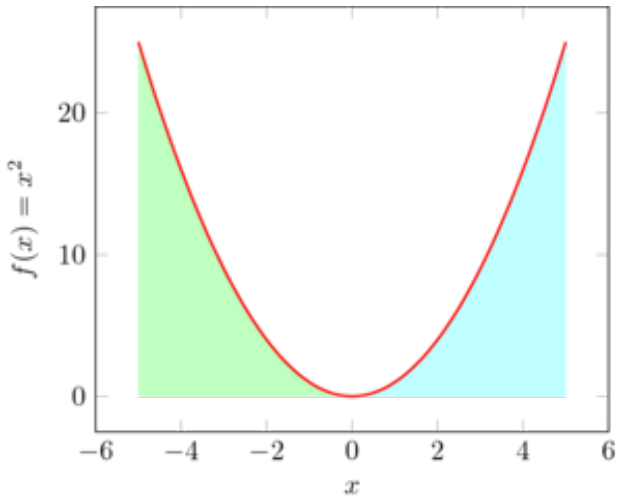
- If $f'(x) > 0$, then $f(x)$ is increasing.
- If $f'(x) < 0$, then $f(x)$ is decreasing.
- If $f'(x) = 0$, then its a critical point unless flat trap or endpoints.

3.2.4 Flat Traps

As said earlier, not all critical points are either global or local maxima or minima. The flat regions might trick one to thin that is a maximum or minimum point. This is where, a small test around the critical point becomes important. Taking a small interval of values around the point, and depending on if $f'(x)$ is increasing or decreasing below and above the point, one could conclude if that critical point was just a flat or extrema (let us refer all critical points which qualify as maximum or minimum as extremum). Refer 1, 2 and practice in same session.

3.2.5 Absolute Minima or Maxima (entire domain)

We already saw a hint to avoid flat traps which we could use to identify the extremum points. Imagine a function like below $f(x) = x^2$. Just by eyeballing we could say, it decreases for interval $x = (-\infty, 0)$ and increases for $x = (0, \infty)$. So the absolute minimum happens at $x = 0$, but how do we prove that mathematically.



We know when the slope is negative, the function is decreasing, and increasing if positive. We could just take a value before and after our critical point, and see if that is the case, to decide if the critical point is minimum or maximum. Let us first find the critical point for $f(x) = x^2$ in the interval $[-\infty, \infty]$. Note, this interval should either be explicitly or implicity defined before we try assessing the critical points. Here, we are able to take entire infinite range because of the ever increasing nature of the function before and after critical point.

$$f'(x) = 2x$$

When $x < 0$, for eg, $x = -2$, then $f'(x) = 2(-2) = -4 < 0$, so its decreasing.
When $x > 0$, for eg, $x = 2$, then $f'(x) = 2(2) = 4 > 0$, so its increasing.

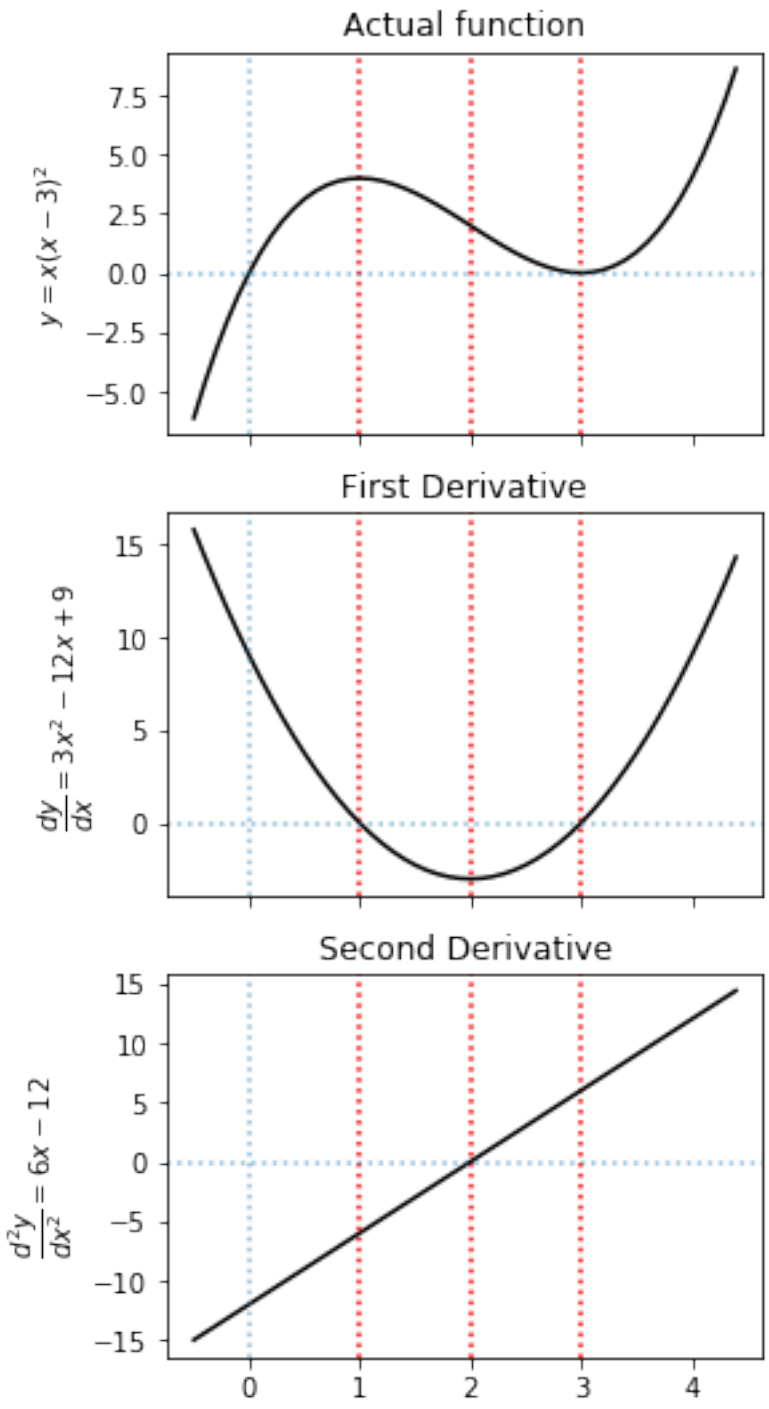
Thus, we infer, before critical point, the given function x^2 is decreasing, and after critical point, the given function x^2 is increasing. Thus the critical point should be a minimum. Since this is in entire domain, this is an absolute minimum point for the function x^2. Refer 1 for another example where the domain is implicit in the function.

If we get more than 1 critical point within the interval, then simply taking the maximum of all the $f(x)$ values at those critical points, will give absolute or global maximum point.

3.2.6 Concavity

It is tedious every time to take values around the first derivative so let us try an easier method of taking second derivative.

Let $f(x) = x(x-3)^2$. The function, and its derivatives will look like below.



By eyeballing, one could see, $f(x)$ reaches a maximum at $x = 1$. Understandly, in 2nd graph, we could observe, that $f'(x)_{x=1} = 0$. Note the 2nd derivative in 3rd graph is giving one more information, that it is negative. That is, $f''(x)_{x=1} < 0$.

$f(x)$ reaches a minimum at $x = 3$. Understandly, $f'(x)_{x=3} = 0$ again. And $f''(x)_{x=3} > 0$, that is, its positive, indicating that the original functino $f(x)$ is increasing.

$f(x)$ reaches an inflection point at $x = 2$. It is a point at which , the first derivative reaches a minimum as seen in 2nd graph. Note at this point, $f''(x)_{x=2} = 0$. In terms of original function, it is a point at which the curve stops being a concave (concave downward) and becomes a conves (or concave upward).

Thus by observing the 2nd derivative, one could conclude about whether a critical point is maximum or minimum or inflection. However, there is a trap, this method is not very rigorous. Refer 1 which explains about the trap.

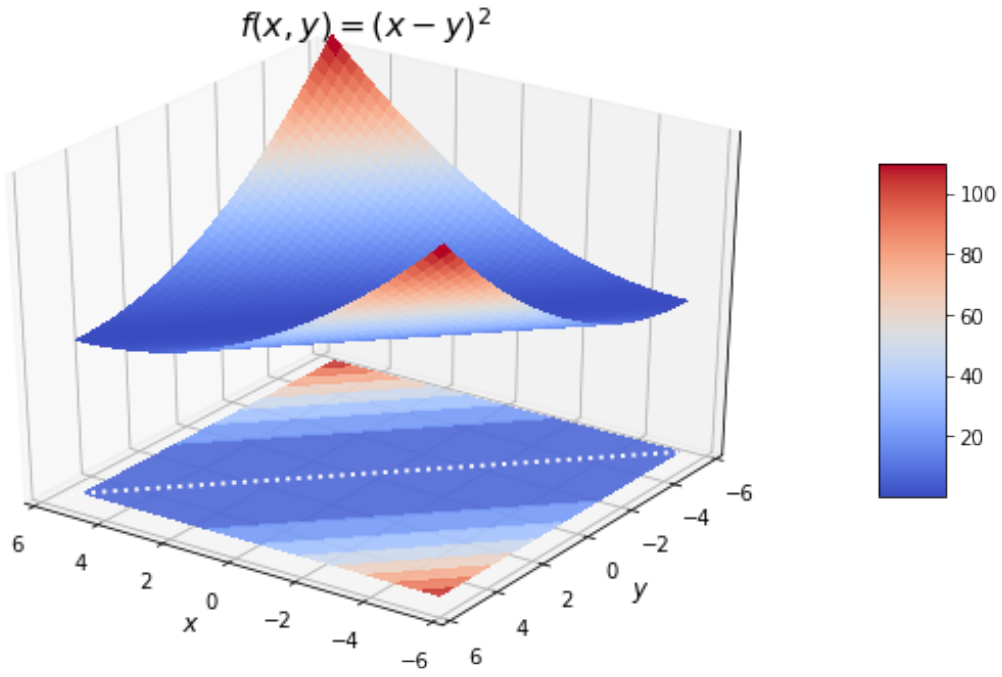
Concavity of given function

- If $f''(x) > 0$ at $f'(x) = 0$, then $f(x)$ is increasing.
- If $f''(x) < 0$ at $f'(x) = 0$, then $f(x)$ is decreasing.
- If $f''(x) = 0$ and $f'(x) \neq 0$, then its an inflection point. But not always.

3.2.7 Surface Plots

The same concepts could also be directly transferred to 3D plots via partial derivatives. For eg, for a function $f(x, y)$, given one of the variables, say x is a constant k, a critical point occurs when $\frac{\partial f(x,y)}{\partial y}\Big|_{x=k} = 0$. Similarly when $y = k$, critical point is expected at $\frac{\partial f(x,y)}{\partial x}\Big|_{y=k} = 0$

Let us consider an example $z = f(x, y) = (y - x)^2$. If we plot the figure, we could already observe that its ever increasing on two directions and has one valley, where the minimum must be occuring. The contour is also shown on XY plane, where one could observe the minimum value occurs along the valley line.



Taking partial derivative with x as constant,

$$\frac{\partial f(x,y)}{\partial y} = \frac{\partial (x-y)^2}{\partial y} = 2(x-y)(-1) = -2(x-y)$$

Assigning it to 0, we get,

$$-2(x-y) = 0 \implies (x-y) = 0 \implies x = y$$

In fact that is what we observed in above diagram. The critical points are in fact a line defined by $x = y$ denoted by dotted white line above. Let us observe what happens for other possibility (we could already observe from graph, that it should result in same output $x = y$), there are no other critical lines. Taking partial derivative with y as constant,

$$\frac{\partial f(x,y)}{\partial x} = \frac{\partial (x-y)^2}{\partial x} = 2(x-y)(1) = 2(x-y)$$

Assigning it to 0, we get,

$$2(x-y) = 0 \implies (x-y) = 0 \implies x = y$$

The same answer. Thus, we are able to mathematically find the critical point. Remember, the first derivative only tells it could be a critical point, not already maximum or minimum, that should be done after wards. In our case, from the graph we got the hint its a minimum, but not yet mathematically.

Second order Trap

However, we cannot just directly interpret second order partial derivatives for $f(x,y)$ like we did for one variable functions. In fact that would be inconclusive. Something more is needed.

Let us try. We shall keep y as constant, say $y = 3$. Taking second order derivative, w.r.t x

$$\frac{\partial^2 f(x,y)}{\partial x} = \frac{\partial^2 (x-y)^2}{\partial x} = \frac{\partial 2(x-y)}{\partial x} = 2 > 0$$

This should mean, our function $f(x,y)$ should be increasing, but if you look at $y = 3$ plane, you could observe that as x is increasing, the function decreased. However, if you look at the plane $y = -3$, $f(x,y)$ is indeed seem to be increasing with x. This is illustrated in Fig 3.1.

Similarly, if we try to keep x as constant, and take partial derivative w.r.t y,

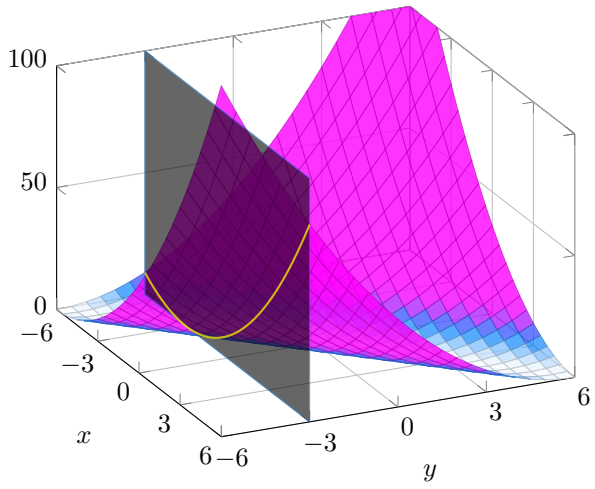
$$\frac{\partial^2 f(x,y)}{\partial y} = \frac{\partial^2 (x-y)^2}{\partial y} = \frac{\partial -2(x-y)}{\partial x} = -2(-1) = 2 > 0$$

Again we see a similar predicament. Observe for both $x = -3$ and $x = 3$ as shown in Fig 3.2. The inconclusivness is because, there is more to surfaces or two variable functions $f(x,y)$ compared to single variable ones. Apart from minium, maximum they also have saddle points. And the possible second order partial derivatives are not just two as we saw, but four as below.

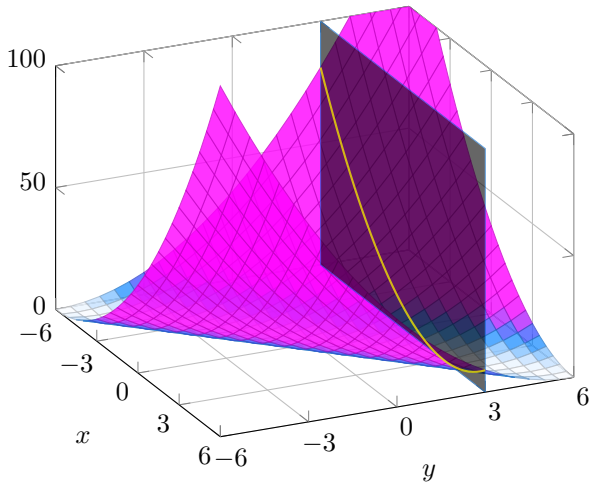
$$f_{xx} = \frac{\partial f^2}{\partial x^2}$$

$$f_{yy} = \frac{\partial f^2}{\partial y^2}$$

$$f_{xy} = \frac{\partial f^2}{\partial x \partial y} = \frac{\partial f^2}{\partial y \partial x}$$

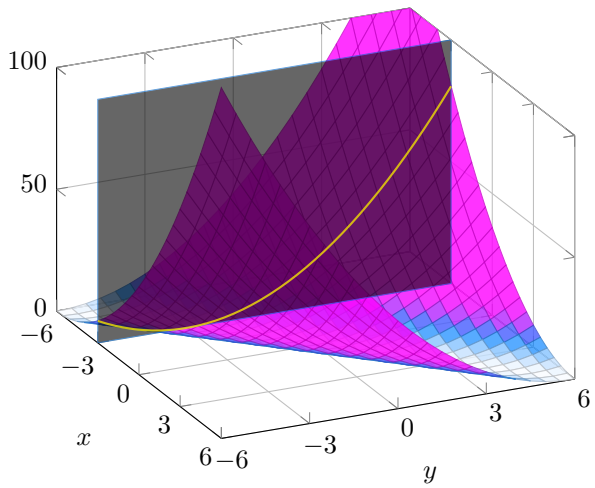


(a) f(x,y) increases with x when $y = -3$

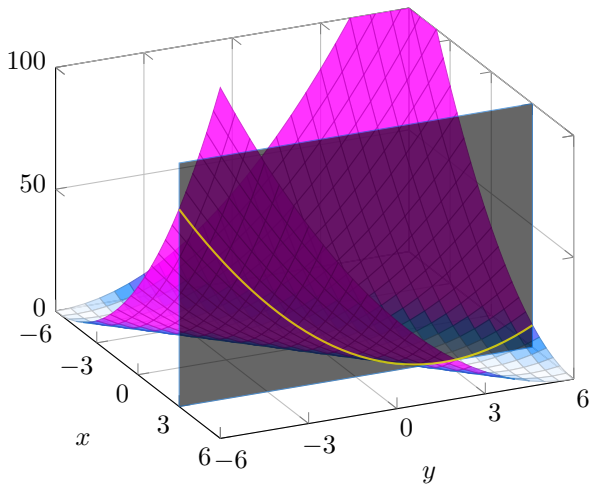


(b) f(x,y) decreases with x when $y = 3$

Fig 1: Inconclusivness from $\frac{\partial^2 f(x,y)}{\partial x^2}\Big|_{y=k}$



(a) f(x,y) increases with y when $x = -3$



(b) f(x,y) decreases with y when $x = 3$

Fig 2: Inconclusivness from $\frac{\partial^2 f(x,y)}{\partial y^2}\Big|_{x=k}$

Thus in case of surfaces, by making a first order partial differentiation w.r.t x and y, what we would get could also be a maximum or minimum or also a saddle point. The method to classify via second order as I just said, is little bit more involved. We will revisit and resume in future if needed, but for a quick dip on working that as well with an example, refer here

Bibliography

[1] J. Devore. *Probability and Statistics for Engineering and the Sciences*. Brooks/Cole CENGAGE Learning, 8th edition, 2011. URL https://fac.ksu.edu.sa/sites/default/files/probability_and_statistics_for_engineering_and_the_sciences.pdf.

[2] J. Frost. Heteroscedasticity in regression analysis. 2017. URL http://statisticsbyjim.com/regression/heteroscedasticity-regression/.

[3] Robert, Elliot, and Dale. *Probability and Statistical Inference*. Pearson, 9th edition, 2015. URL http://www.nylxs.com/docs/thesis/sources/Probability%20and%20Statistical%20Inference%209ed%20%5B2015%5D.pdf.